**ĐẠI HỌC UEH**

**KHOA CÔNG NGHỆ THÔNG TIN KINH DOANH**

# **MỤC LỤC**

[**MỤC LỤC** ii](#_Toc198724867)

[**DANH MỤC BẢNG BIỂU** iv](#_Toc198724868)

[**DANH MỤC HÌNH ẢNH** v](#_Toc198724869)

[**CHƯƠNG 1: GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU VÀ ĐỀ TÀI** 1](#_Toc198724870)

[**CHƯƠNG 2: TỔNG QUAN VỀ CHƯƠNG TRÌNH SỬ DỤNG VÀ CÁC PHƯƠNG PHÁP SỬ DỤNG** 2](#_Toc198724871)

[**2.1. Giới thiệu tập dữ liệu** 2](#_Toc198724872)

[**2.2. Mô tả các biến** 3](#_Toc198724873)

[**2.3. Khám phá dữ liệu ban đầu (EDA)** 7](#_Toc198724874)

[**2.3.1. EDA Customer table** 7](#_Toc198724875)

[**2.3.2. EDA Ticket table** 13](#_Toc198724876)

[**2.4. Các bước tiền xử lý dữ liệu** 20](#_Toc198724877)

[**2.4.1. Xử lý bảng TICKET** 20](#_Toc198724878)

[**CHƯƠNG 3: ỨNG DỤNG CÁC PHƯƠNG PHÁP VÀO BÀI TOÁN KINH DOANH** 27](#_Toc198724879)

[**3.2. Lựa chọn phương pháp** 27](#_Toc198724880)

[**3.4. Quy trình huấn luyện mô hình** 29](#_Toc198724881)

[**CHƯƠNG 4: KẾT QUẢ MÔ HÌNH** 31](#_Toc198724882)

[**4.1. Tiền xử lý dữ liệu** 31](#_Toc198724883)

[**4.2. Kết quả các mô hình** 31](#_Toc198724884)

[**4.3. Phân tích và so sánh** 36](#_Toc198724885)

[**4.3.1. Phân tích kết quả từ thuật toán KMean** 36](#_Toc198724886)

[**4.3.2. Phân tích kết quả từ thuật toán DBSCAN** 37](#_Toc198724887)

[**4.4. Dashboard tổng hợp kết quả:** 38](#_Toc198724888)

[**4.3.1. Overview Dashboard** 39](#_Toc198724889)

[**4.3.2. Customer Dashboard** 49](#_Toc198724890)

[**4.3.3. Scheduling Dashboard** 51](#_Toc198724891)

[**PHỤ LỤC** vi](#_Toc198724892)

[**Phụ lục 1: Khám phá dữ liệu ban đầu (EDA)** vi](#_Toc198724893)

[**Phụ lục 2: Tiền xử lý dữ liệu** vi](#_Toc198724894)

# **DANH MỤC BẢNG BIỂU**

[Bảng 2.1: Mô tả các biến 2](#_Toc198724491)

# **DANH MỤC HÌNH ẢNH**

[Hình 2.1: Biểu đồ số lượng giá trị thiếu theo cột 6](#_Toc198724298)

[Hình 2.2: Biểu đồ phân phối của biến Gender 7](#_Toc198724299)

[Hình 2.3: Biểu đồ phân phối của biến Job 8](#_Toc198724300)

[Hình 2.4: Biểu đồ phân phối của biến Industry 9](#_Toc198724301)

[Hình 2.5: Biểu đồ phân phối nghề nghiệp theo lĩnh vực ngành 10](#_Toc198724302)

[Hình 2.6: Biểu đồ phân phối nghề nghiệp theo giới tính 11](#_Toc198724303)

[Hình 2.7: Thống kê mô tả bảng Ticket 12](#_Toc198724304)

[Hình 2.8: Số lượng giá trị thiếu của từng biến 13](#_Toc198724305)

[Hình 2.9: Ma trận tương quan giữa các biến số (Numerical) 14](#_Toc198724306)

[Hình 2.10: Biểu đồ phân phối của biến Ticket price 15](#_Toc198724307)

[Hình 2.11: Biểu đồ phân phối của biến Total 16](#_Toc198724308)

[Hình 2.12: Biểu đồ tần suất sử dụng phòng chiếu 17](#_Toc198724309)

[Hình 2.13: Biểu đồ phân phối tần suất chiếu phim 18](#_Toc198724310)

[Hình 2.14: Phân phối tần suất hoạt động của nhân viên thu ngân 19](#_Toc198724311)

[Hình 3.1: Minh họa cho mô hình phân cụm 27](#_Toc198724294)

[Hình 4.1: Biểu đồ Elbow – Chọn số cụm tối ưu 31](#_Toc198724288)

[Hình 4.2: Mô tả kết quả phân cụm KMeans 33](#_Toc198724289)

[Hình 4.3: Mô tả kết quả phân cụm DBSCAN 34](#_Toc198724290)

[Hình 4.4: Dashboard Overview - Tổng quát về tình trạng doanh nghiệp 38](#_Toc198724291)

[Hình 4.5: Customer Dashboard - Chân dung khách hàng 48](#_Toc198724292)

[Hình 4.6: Scheduling Dashboard - Xây dựng lịch chiếu 50](#_Toc198724293)

# 

# **CHƯƠNG 1: GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU VÀ ĐỀ TÀI**

Khoa học dữ liệu (Data Science) là một lĩnh vực liên ngành tiên tiến, kết hợp giữa toán học, thống kê, khoa học máy tính và kiến thức chuyên ngành để trích xuất thông tin và hiểu biết từ dữ liệu phức tạp. Trong bối cảnh cuộc cách mạng công nghiệp 4.0 và sự bùng nổ của dữ liệu lớn (Big Data), khoa học dữ liệu đóng vai trò then chốt trong việc chuyển hóa dữ liệu thô thành tri thức có giá trị, hỗ trợ quá trình ra quyết định trong nhiều lĩnh vực như kinh doanh, y tế, tài chính và khoa học xã hội. Các nhà khoa học dữ liệu sử dụng các kỹ thuật như học máy, khai phá dữ liệu và trực quan hóa dữ liệu để phát hiện các mẫu và xu hướng, từ đó đưa ra các dự đoán và giải pháp chiến lược.

Trong bối cảnh ngành công nghiệp giải trí ngày càng phát triển và cạnh tranh khốc liệt, việc ứng dụng Business Intelligence (BI) vào hoạt động quản lý và vận hành rạp chiếu phim trở nên thiết yếu. Đề tài "Phân tích dữ liệu rạp chiếu phim" nhằm mục tiêu khai thác và phân tích các nguồn dữ liệu như lịch chiếu, doanh thu bán vé, số lượng khán giả để cung cấp những thông tin chi tiết hỗ trợ quá trình ra quyết định. Thông qua việc áp dụng các công cụ BI, nhóm nghiên cứu sẽ xây dựng hệ thống báo cáo và trực quan hóa dữ liệu, giúp ban quản lý rạp phim nắm bắt kịp thời xu hướng thị trường, tối ưu hóa lịch chiếu, quản lý hiệu quả nguồn lực và nâng cao trải nghiệm khách hàng. Đề tài hướng đến việc phần nâng cao hiệu quả kinh doanh của rạp chiếu phim và mở ra hướng tiếp cận mới trong việc ứng dụng BI vào lĩnh vực giải trí.

# **CHƯƠNG 2: TỔNG QUAN VỀ CHƯƠNG TRÌNH SỬ DỤNG VÀ CÁC PHƯƠNG PHÁP SỬ DỤNG**

Trong quá trình thực hiện đề tài, nhóm chúng em sử dụng kết hợp Excel và Python để xử lý, phân tích dữ liệu từ hệ thống rạp chiếu phim. Excel được ứng dụng trong các bước tiền xử lý cơ bản như làm sạch dữ liệu, chuẩn hóa định dạng và kiểm tra tính đầy đủ của thông tin. Song song đó, Python với các thư viện chuyên biệt như Pandas, Scikit-learn và Seaborn được triển khai nhằm thực hiện các phân tích nâng cao, đặc biệt là kỹ thuật phân cụm (clustering) để nhận diện nhóm khách hàng theo hành vi mua vé và sở thích xem phim. Kết quả phân tích sau đó được trực quan hóa bằng Power BI, giúp thể hiện trực quan các xu hướng, mô hình và mối liên hệ quan trọng trong dữ liệu. Việc tích hợp các công cụ này không chỉ nâng cao tính chính xác và hiệu quả trong phân tích mà còn hỗ trợ trực tiếp cho việc ra quyết định chiến lược trong quản lý và vận hành rạp chiếu phim.

## **2.1. Giới thiệu tập dữ liệu**

Tập dữ liệu sử dụng được thu thập thủ công từ hệ thống bán vé và quản lý nội dung phim tại rạp CGV Đà Nẵng trong giai đoạn tháng 5/2019. Dữ liệu phản ánh hoạt động thực tế của khách hàng, các giao dịch mua vé cũng như thông tin mô tả về từng bộ phim được chiếu tại rạp trong giai đoạn khảo sát. Việc thu thập và xử lý dữ liệu được thực hiện nhằm phục vụ mục tiêu phân tích hành vi người dùng và hỗ trợ ra quyết định trong lĩnh vực vận hành rạp chiếu phim.

Dữ liệu được lưu trữ dưới dạng hai tệp Excel riêng biệt. Trong đó, tệp processed\_ticket\_cleaned là nguồn dữ liệu chính phục vụ cho cả quá trình phân tích và trực quan hóa. Tệp này bao gồm hai bảng thành phần:

* CUSTOMER: ghi nhận thông tin nhân khẩu học của khách hàng,
* TICKET: phản ánh chi tiết các giao dịch bán vé tại rạp trong suốt thời gian khảo sát.

Bên cạnh đó, tệp MOVIE\_TABLE cung cấp thông tin mô tả về các bộ phim được trình chiếu, bao gồm các biến như tên phim, thể loại, thời lượng, quốc gia sản xuất và phân loại độ tuổi. Dữ liệu từ MOVIE\_TABLE được sử dụng nhằm chuẩn hóa thông tin phim và phục vụ đối chiếu mã phim trong quá trình kết nối với bảng giao dịch. Việc chuẩn hóa thông tin phim từ bảng MOVIE\_TABLE giúp đảm bảo tính nhất quán và hỗ trợ đối chiếu mã phim (movie\_id) trong quá trình kết nối giữa các bảng.

Các bảng dữ liệu được liên kết logic với nhau thông qua các biến định danh khóa như customerid, orderid, và movie\_id, cho phép thực hiện thao tác nối bảng (join) một cách thuận tiện khi xây dựng báo cáo và mô hình phân tích.

Xét về quy mô:

* Bảng CUSTOMER bao gồm 4.479 bản ghi và 7 biến đặc trưng cho từng khách hàng
* Bảng TICKET bao gồm 35.508 bản ghi với 18 biến, phản ánh chi tiết từng vé được bán
* Bảng MOVIE\_TABLE chứa 19 bản ghi và 8 biến, đại diện cho danh mục phim đã được chuẩn hóa

Mỗi bản ghi trong các bảng tương ứng đại diện cho một đơn vị quan sát độc lập (ví dụ: một khách hàng, một giao dịch, hoặc một bộ phim), trong khi các biến phản ánh các đặc điểm liên quan đến đối tượng đó.

## **2.2. Mô tả các biến**

Bảng 2.1: Mô tả các biến

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên biến** | **Kiểu dữ liệu** | **Mô tả** | **Nguồn dữ liệu** |
| 1 | customerid | Identifier | Mã định danh của từng khách hàng | processed\_ticket\_cleaned – CUSTOMER |
| 2 | DOB | Time-based (Date) | Ngày sinh của khách hàng | processed\_ticket\_cleaned – CUSTOMER |
| 3 | gender | Categorical | Giới tính của khách hàng (Nam/Nữ) | processed\_ticket\_cleaned – CUSTOMER |
| 4 | address | Text | Địa chỉ nơi sinh sống hoặc nơi đăng ký của khách hàng | processed\_ticket\_cleaned – CUSTOMER |
| 5 | Website | Categorical | Kênh tiếp cận hoặc mã nhận diện người dùng qua hệ thống bán vé | processed\_ticket\_cleaned – CUSTOMER |
| 6 | job | Categorical | Nghề nghiệp của khách hàng (student, white collar, specialist, blue collar, teenager) | processed\_ticket\_cleaned – CUSTOMER |
| 7 | industry | Categorical | Ngành/lĩnh vực làm việc của khách hàng (health service, finance, computer, construction,...) | processed\_ticket\_cleaned – CUSTOMER |
| 8 | orderid | Identifier | Mã đơn hàng | processed\_ticket\_cleaned – TICKET |
| 9 | cashier | Categorical | Mã nhân viên bán vé | processed\_ticket\_cleaned – TICKET |
| 10 | saledate | Time-based (Datetime) | Thời điểm khách hàng mua vé tại quầy | processed\_ticket\_cleaned – TICKET |
| 11 | customerid | Identifier | Mã định danh khách hàng | processed\_ticket\_cleaned – TICKET |
| 12 | date | Time-based (Date) | Ngày chiếu phim | processed\_ticket\_cleaned – TICKET |
| 13 | time | Time-based (Time) | Giờ bắt đầu suất chiếu | processed\_ticket\_cleaned – TICKET |
| 14 | room | Categorical | Phòng chiếu phim (1, 2, 3, 4, 5, 6) | processed\_ticket\_cleaned – TICKET |
| 15 | film | Text | Tên phim ghi nhận theo giao dịch | processed\_ticket\_cleaned – TICKET |
| 16 | ticketcode | Identifier | Mã vé xem phim | processed\_ticket\_cleaned – TICKET |
| 17 | slot | Categorical | Mã ghế trong phòng chiếu | processed\_ticket\_cleaned – TICKET |
| 18 | ticket price | Numerical | Giá cho một vé xem phim | processed\_ticket\_cleaned – TICKET |
| 19 | total | Numerical | Tổng tiền cho cả đơn hàng (có thể gồm nhiều vé) | processed\_ticket\_cleaned – TICKET |
| 20 | popcorn | Boolean (0/1) | Khách có mua bắp nước hay không (Có, Không) | processed\_ticket\_cleaned – TICKET |
| 21 | slot type | Categorical | Loại ghế (đơn, đôi) | processed\_ticket\_cleaned – TICKET |
| 22 | ticket type | Categorical | Loại vé ghi nhận trong giao dịch (Thành viên, Không là thành viên) | processed\_ticket\_cleaned – TICKET |
| 23 | discount | Numerical | Mức giảm giá áp dụng (nếu có) | processed\_ticket\_cleaned – TICKET |
| 24 | have\_discount | Boolean (0/1) | Vé có được giảm giá hay không (Có, Không) | processed\_ticket\_cleaned – TICKET |
| 25 | movie\_id | Identifier | Mã định danh duy nhất của phim trong hệ thống | MOVIE\_TABLE |
| 26 | title | Text | Tên bộ phim | MOVIE\_TABLE |
| 27 | country | Categorical | Quốc gia phát hành hoặc sản xuất phim | MOVIE\_TABLE |
| 28 | languages | Categorical | Ngôn ngữ sử dụng trong phim | MOVIE\_TABLE |
| 29 | release\_year | Time-based (Year) | Năm phát hành phim | MOVIE\_TABLE |
| 30 | rating\_viet | Categorical | Phân loại độ tuổi theo tiêu chuẩn Việt Nam (P, C13, C18...) | MOVIE\_TABLE |
| 31 | duration | Numerical | Thời lượng phim tính theo phút | MOVIE\_TABLE |
| 32 | listed\_in | Text | Chuỗi thể loại phim, gồm nhiều nhãn phân cách bằng dấu phẩy | MOVIE\_TABLE |

*Nguồn: Nhóm tác giả tự tổng hợp*

## **2.3. Khám phá dữ liệu ban đầu (EDA)**

### **2.3.1. EDA Customer table**

* Kiểm tra missing value

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Hình chữ nhật

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.1: Biểu đồ số lượng giá trị thiếu theo cột

Nguồn: Nhóm tác giả tính toán

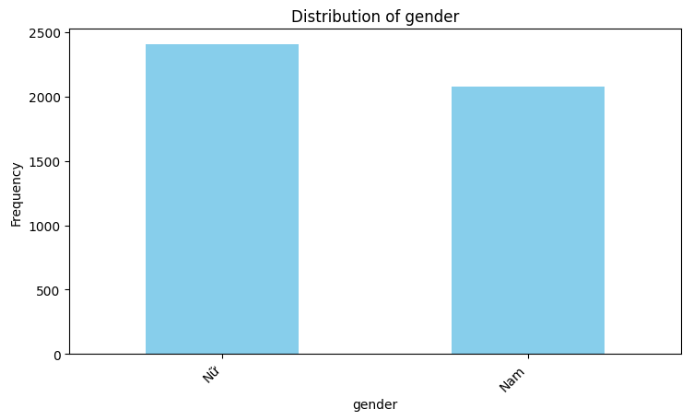
Nhận xét : Qua quan sát, có thể thấy rằng hầu hết các cột trong tập dữ liệu đều có mức độ hoàn chỉnh khá cao. Cụ thể, các cột như customerid, DOB, gender và job hầu như không có giá trị thiếu, điều này phản ánh dữ liệu ở các trường này được thu thập đầy đủ và có thể tin cậy cho các phân tích định danh, nhân khẩu học hoặc nghề nghiệp.

Tuy nhiên, một số cột khác lại có sự thiếu hụt nhất định cần được chú ý. Cột address xuất hiện một lượng nhỏ giá trị thiếu (dưới 10), tuy không đáng kể so với quy mô toàn bộ tập dữ liệu, nhưng vẫn cần được cân nhắc xử lý nhằm tránh gây sai lệch khi phân tích liên quan đến vị trí địa lý của khách hàng.

Đáng lưu ý hơn, cột Website có số lượng giá trị thiếu lên đến hơn 500, chiếm một tỷ lệ đáng kể trong tập dữ liệu. Cần xem xét việc xử lý thiếu hụt này bằng cách loại bỏ, thay thế hoặc suy diễn dựa trên các trường liên quan nếu phù hợp.

Nghiêm trọng nhất là cột industry, với hơn 1.100 giá trị bị thiếu – đây là tỷ lệ thiếu rất lớn và có khả năng ảnh hưởng trực tiếp đến các phân tích theo lĩnh vực hoạt động của khách hàng. Thiếu hụt dữ liệu ở cột này không chỉ làm giảm khả năng phân loại, mà còn gây khó khăn trong việc xây dựng các mô hình dự đoán hoặc đề xuất phù hợp theo ngành nghề. Do đó, việc xử lý và cân nhắc kỹ lưỡng đối với trường này là cần thiết trước khi tiến hành các bước phân tích sâu hơn.

* Phân phối của biến Gender



Hình 2.2: Biểu đồ phân phối của biến Gender

Nguồn: Nhóm tác giả tính toán

Nhận xét :Trong quá trình phân tích nhân khẩu học của khách hàng, yếu tố giới tính là một trong những biến quan trọng giúp hiểu rõ hơn về đặc điểm và hành vi tiêu dùng của đối tượng mục tiêu. Dữ liệu cho thấy phân bố giới tính trong tập khách hàng không hoàn toàn cân bằng, với số lượng khách hàng nữ chiếm ưu thế so với nam giới. Cụ thể, có khoảng 2.400 khách hàng là nữ, trong khi số lượng khách hàng nam chỉ vào khoảng 2.100.

* Phân phối của biến job

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, hàng

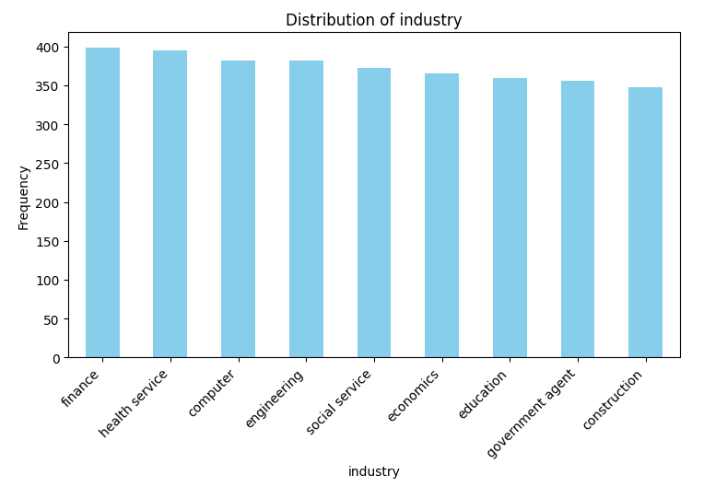
Nội dung do AI tạo ra có thể không chính xác.

Hình 2.3: Biểu đồ phân phối của biến Job

Nguồn: Nhóm tác giả tính toán

Nhận xét : Phân tích phân bố nghề nghiệp của khách hàng cho thấy sự không đồng đều rõ rệt giữa các nhóm, phản ánh đặc điểm nhân khẩu học đặc thù của rạp chiếu phim này. Cụ thể, nhóm khách hàng là "student" (sinh viên) chiếm tỷ trọng lớn nhất, vượt trội so với tất cả các nhóm nghề nghiệp khác. Điều này cho thấy sinh viên là nhóm đối tượng quan trọng và có sức thu hút cao đối với các hoạt động giải trí như xem phim, có thể do sự linh hoạt về thời gian và nhu cầu giải trí cao của nhóm tuổi này. Đứng thứ hai là nhóm "teenager" (thiếu niên), cũng chiếm một tỷ lệ đáng kể trong tổng số khách hàng. Sự hiện diện lớn của nhóm tuổi trẻ như sinh viên và thiếu niên cho thấy rạp chiếu phim đang phục vụ chủ yếu cho các đối tượng trẻ, năng động và có xu hướng tiêu dùng theo xu hướng mới. Trong khi đó, các nhóm nghề nghiệp khác như "specialist" (chuyên gia), "white collar" (nhân viên văn phòng) và "blue collar" (lao động phổ thông) có số lượng khách hàng thấp hơn đáng kể và phân bố khá tương đồng với nhau. Tổng thể, dữ liệu phản ánh sự phân bố không đồng đều giữa các nhóm nghề nghiệp, với sự áp đảo của sinh viên và thiếu niên. Thông tin này là cơ sở quan trọng để rạp chiếu phim điều chỉnh chiến lược marketing, lựa chọn nội dung phim, thời gian chiếu và các chương trình ưu đãi sao cho phù hợp với nhu cầu và thói quen tiêu dùng của nhóm khách hàng chủ lực là giới trẻ.

* Phân phối của biến industry



Hình 2.4: Biểu đồ phân phối của biến Industry

Nguồn: Nhóm tác giả tính toán

Nhận xét : Phân tích phân bố khách hàng theo ngành nghề cho thấy mức độ tiếp cận rộng rãi và tính phổ quát của rạp chiếu phim đối với nhiều nhóm ngành khác nhau trong xã hội. Biểu đồ minh họa cho thấy sự phân bố tương đối đồng đều giữa các ngành, khi không có ngành nào chiếm ưu thế tuyệt đối về số lượng khách hàng. Các cột đại diện cho từng ngành có chiều cao khá đồng đều, điều này phản ánh rằng nhu cầu xem phim là một hình thức giải trí phổ biến, không giới hạn trong một nhóm ngành cụ thể. Tuy nhiên, một số sự khác biệt nhỏ vẫn có thể được nhận thấy. Ngành "finance" (tài chính) và "health service" (dịch vụ y tế) có số lượng khách hàng nhỉnh hơn so với các ngành còn lại. Dù mức chênh lệch không quá lớn, điều này có thể gợi ý rằng các đối tượng làm việc trong môi trường chuyên môn cao hoặc có thu nhập ổn định có xu hướng dành thời gian cho các hoạt động giải trí như xem phim thường xuyên hơn. Ngược lại, ngành "construction" (xây dựng) có số lượng khách hàng thấp nhất trong số các ngành được khảo sát. Nguyên nhân có thể xuất phát từ đặc điểm công việc đòi hỏi thời gian linh hoạt hoặc ít tiếp cận với các kênh giải trí truyền thống.

* Phân phối của biến Job và Industry

Ảnh có chứa văn bản, ảnh chụp màn hình, Sơ đồ, Nhiều màu sắc

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.5: Biểu đồ phân phối nghề nghiệp theo lĩnh vực ngành

Nguồn: Nhóm tác giả tính toán

Nhận xét :Nhóm sinh viên (student) tiếp tục chiếm tỷ trọng vượt trội so với các nhóm nghề nghiệp còn lại ở hầu hết các ngành, đặc biệt là các ngành social service, finance và health service, cho thấy rạp chiếu phim có sức hút lớn đối với nhóm khách hàng trẻ đang trong quá trình học tập ở nhiều lĩnh vực khác nhau. Ngược lại, nhóm teenager (thiếu niên) gần như không xuất hiện trong biểu đồ, điều này có thể phản ánh sự giới hạn trong dữ liệu. Các nhóm nghề khác như blue collar, specialist và white collar có phân bố tương đối đồng đều, tuy nhiên vẫn thấp hơn đáng kể so với nhóm sinh viên. Đáng chú ý, nhóm specialist có sự nổi bật trong các ngành như engineering và finance, trong khi nhóm white collar thể hiện sự quan tâm rõ rệt đến ngành government agent và engineering. Nhóm blue collar cũng có sự hiện diện đáng kể trong ngành education, health service và engineering. Từ biểu đồ này, có thể thấy rằng sự đa dạng ngành nghề của khách hàng phần lớn tập trung trong nhóm sinh viên, trong khi các nhóm nghề nghiệp khác có xu hướng tập trung hơn vào một số lĩnh vực chuyên biệt. Điều này cung cấp cơ sở quan trọng để rạp chiếu phim có thể thiết kế các chiến lược tiếp thị và nội dung phim phù hợp với từng phân khúc nghề nghiệp và ngành nghề cụ thể, đồng thời tận dụng tiềm năng từ nhóm khách hàng sinh viên để mở rộng ảnh hưởng trong cộng đồng.

* Phân phối của biến Gender và Job

Ảnh có chứa ảnh chụp màn hình, văn bản, biểu đồ, Sơ đồ

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.6: Biểu đồ phân phối nghề nghiệp theo giới tính

Nguồn: Nhóm tác giả tính toán

Nhận xét :Nhóm sinh viên là lực lượng khách hàng chiếm ưu thế ở cả hai giới, cho thấy đây là đối tượng trung tâm trong tập khách hàng hiện tại. Tuy nhiên, số lượng sinh viên nữ vượt trội so với sinh viên nam, phản ánh xu hướng tiêu dùng giải trí cao hơn của nữ giới trong độ tuổi học sinh – sinh viên. Bên cạnh đó, nhóm thiếu niên (teenager) là nhóm lớn thứ hai ở nữ, nhưng lại có số lượng thấp hơn đáng kể ở nam, cho thấy sự khác biệt rõ nét trong thói quen giải trí giữa hai giới ở độ tuổi vị thành niên. Đối với nam giới, sau nhóm sinh viên là sự phân bổ tương đối đồng đều giữa các nhóm nghề nghiệp trưởng thành như công nhân (blue collar), chuyên gia (specialist) và dân văn phòng (white collar), trong khi nhóm thiếu niên ít xuất hiện hơn. Ngoài ra, nhóm white collar có tỷ lệ khách hàng nam và nữ khá tương đồng, cho thấy mức độ tiêu dùng giải trí cân bằng giữa hai giới trong nhóm nghề nghiệp này. Những xu hướng này không chỉ phản ánh đặc điểm nhân khẩu học của khách hàng, mà còn cung cấp thông tin quan trọng cho việc xây dựng các chương trình tiếp thị và nội dung phù hợp với từng nhóm giới tính và nghề nghiệp cụ thể, từ đó nâng cao hiệu quả thu hút và giữ chân khán giả.

### **2.3.2. EDA Ticket table**

* Thống kê mô tả

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ, hàng

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.7: Thống kê mô tả bảng Ticket

Nguồn: Nhóm tác giả tính toán

* Kiểm tra missing value

Ảnh có chứa văn bản, ảnh chụp màn hình, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.8: Số lượng giá trị thiếu của từng biến

Nguồn: Nhóm tác giả tính toán

Nhận xét : Dữ liệu được đánh giá là khá đầy đủ và chất lượng cao, khi hầu hết các cột như orderid, cashier, customerid, date, time, room, film, ticketcode, slot, ticket price, total, slot type, ticket type, discount và have\_discount đều không có giá trị thiếu. Việc không tồn tại giá trị null ở những cột quan trọng này giúp đảm bảo tính toàn vẹn của dữ liệu, từ đó hỗ trợ tốt cho các phân tích và mô hình hóa sau này. Tuy nhiên vàn tồn tại hai cột có giá trị thiếu là saledate và popcorn, với cùng số lượng thiếu là 96. Việc hai cột này có số lượng thiếu giống nhau đặt ra nghi vấn về mối quan hệ tiềm ẩn giữa ngày bán và thông tin mua bắp nước — có thể là do một nhóm giao dịch không ghi nhận thông tin thời gian bán hoặc không liên quan đến sản phẩm bắp nước. Do đó, cần thực hiện kiểm tra sâu hơn để xác định nguyên nhân gốc rễ, từ đó có chiến lược xử lý giá trị thiếu phù hợp như loại bỏ, điền giá trị thay thế hoặc phân tích riêng biệt nhóm dữ liệu bị thiếu.

* Tương quan giữa các biến số

Ảnh có chứa văn bản, ảnh chụp màn hình, hình vuông, Hình chữ nhật

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.9: Ma trận tương quan giữa các biến số (Numerical)

Nguồn: Nhóm tác giả tính toán

Nhận xét : Phân tích ma trận tương quan cho thấy phần lớn các đặc trưng số trong tập dữ liệu không có mối quan hệ tuyến tính đáng kể với nhau, thể hiện qua hệ số tương quan gần bằng 0 giữa hầu hết các cặp biến. Điều này phản ánh rằng các đặc trưng này hoạt động tương đối độc lập và không có sự phụ thuộc tuyến tính rõ rệt. Trong số đó, cặp "room" và "ticket price" có mối tương quan âm nhẹ với hệ số -0.17, cho thấy có thể tồn tại một xu hướng nhỏ mà giá vé giảm nhẹ ở một số phòng chiếu nhất định. Tuy nhiên, mối liên hệ này còn khá yếu và cần được kiểm tra sâu hơn bằng cách phân tích thêm thông tin về từng phòng chiếu cụ thể. Ngoài ra, các cặp còn lại như "room" với "ticketcode", "total", "discount", hay "ticket price" với "total", "discount", đều có hệ số tương quan rất nhỏ (trong khoảng -0.03 đến 0.01), chứng tỏ không có mối quan hệ tuyến tính đáng kể nào. Một điểm tích cực là không có hiện tượng đa cộng tuyến trong dữ liệu, vì không có cặp biến nào có hệ số tương quan tuyệt đối gần bằng 1. Điều này giúp đảm bảo tính ổn định và tin cậy khi xây dựng các mô hình thống kê hoặc học máy sử dụng những đặc trưng này.

* Phân phối của biến Ticket price

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, hàng

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.10: Biểu đồ phân phối của biến Ticket price

Nguồn: Nhóm tác giả tính toán

**Nhận xét :** Phân tích phân bố giá vé cho thấy một đặc điểm lệch phải rõ rệt, phản ánh sự chênh lệch lớn giữa số lượng các giao dịch ở mức giá thấp và cao. Cụ thể, phần lớn các giao dịch tập trung dày đặc trong khoảng giá từ 40,000 đến 50,000 đồng, với một đỉnh rất cao tại vùng này, cho thấy đây là mức giá vé phổ biến nhất và được khách hàng lựa chọn nhiều nhất. Tuy nhiên, bên cạnh đó vẫn tồn tại một số lượng nhỏ các giao dịch với giá vé cao hơn đáng kể, trải dài từ khoảng 70,000 đến 160,000 đồng. Điều này tạo ra một cái đuôi dài về bên phải trong biểu đồ phân bố, đặc trưng cho một phân bố lệch phải. Đáng chú ý, một số đỉnh nhỏ hơn xuất hiện tại khoảng giá 80,000 và 95,000 đồng. Nhìn chung, cấu trúc phân bố này cho thấy sự tồn tại của các phân khúc giá khác nhau trong hoạt động bán vé, đồng thời phản ánh thói quen tiêu dùng chủ yếu tập trung ở mức giá vé phổ thông.

* Phân phối của biến Total

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Nội dung do AI tạo ra có thể không chính xác.

Hình 2.11: Biểu đồ phân phối của biến Total

Nguồn: Nhóm tác giả tính toán

Nhận xét : Phân tích phân bố tổng số tiền thanh toán cho thấy một đặc điểm lệch phải rõ rệt, tương tự như phân bố giá vé. Cụ thể, phần lớn các giao dịch tập trung ở mức giá trị thấp, gần với 0, với một đỉnh rất cao tại khu vực này, phản ánh rằng đa số khách hàng thực hiện các giao dịch có tổng giá trị nhỏ—có thể chỉ bao gồm một vé xem phim mà không kèm theo các dịch vụ bổ sung như đồ ăn hay nước uống. Tuy nhiên, biểu đồ cũng cho thấy một đuôi dài về phía bên phải, với một số giao dịch có tổng số tiền thanh toán cao vượt trội, thậm chí lên đến gần 2 triệu đồng. Điều này cho thấy sự hiện diện của những khách hàng có mức chi tiêu rất lớn trong một lần giao dịch, có thể bao gồm việc mua nhiều vé hoặc sử dụng thêm các dịch vụ như bắp nước, combo gia đình. Sự chênh lệch lớn giữa các giao dịch có giá trị thấp và cao phản ánh sự đa dạng trong hành vi tiêu dùng của khách hàng, đồng thời cũng mở ra cơ hội để cá nhân hóa dịch vụ và xây dựng các gói sản phẩm phù hợp với từng nhóm khách hàng cụ thể.

* Tần suất phòng chiếu

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, Phông chữ

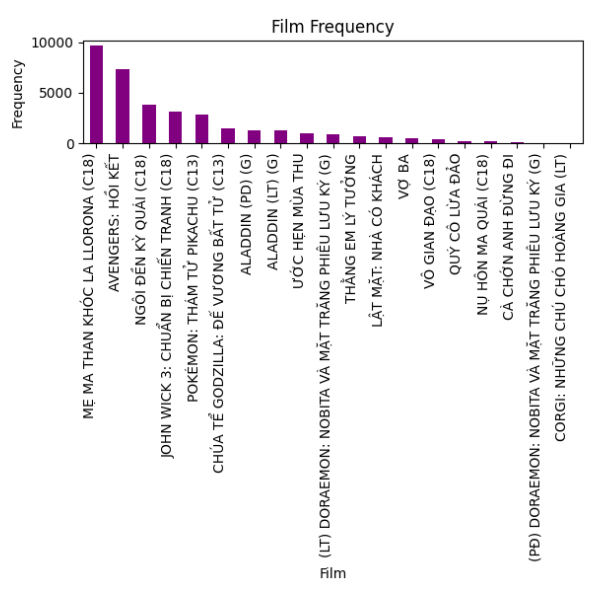
Nội dung do AI tạo ra có thể không chính xác.

Hình 2.12: Biểu đồ tần suất sử dụng phòng chiếu

Nguồn: Nhóm tác giả tính toán

Nhận xét: Biểu đồ tần suất phòng chiếu cho thấy sự phân bổ khá đồng đều giữa các phòng, tuy nhiên phòng chiếu số 4 có tần suất sử dụng cao nhất, với hơn 10.000 lượt, cao hơn so với các phòng còn lại. Ba phòng chiếu còn lại (phòng 1, 2 và 3) có số lượng lượt sử dụng dao động xung quanh mức 8.000 đến 9.000 lượt, với sự khác biệt không quá lớn. Điều này cho thấy các phòng chiếu đều được khai thác tương đối đồng đều, tuy nhiên có thể phòng 4 có sức chứa lớn hơn, được ưu tiên cho các suất chiếu có lượng khách cao, hoặc được bố trí chiếu các phim hút khách.

* Tần suất phim

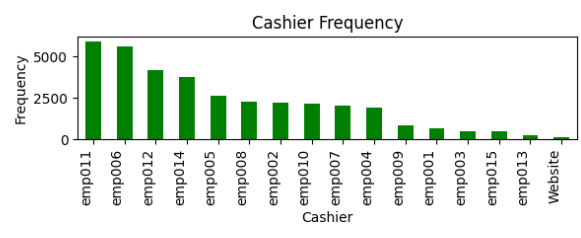


Hình 2.13: Biểu đồ phân phối tần suất chiếu phim

Nguồn: Nhóm tác giả tính toán

Nhận xét: Biểu đồ tần suất phim cho thấy sự phân hóa rõ rệt trong mức độ phổ biến của các bộ phim được chiếu. Dẫn đầu là phim "Mẹ Ma Thần Khóc La Llorona (C18)" với tần suất gần 10.000 lượt, cho thấy đây là bộ phim thu hút đông đảo khán giả nhất. Theo sau là "Avengers: Hồi Kết" và "Ngôi Đền Kỳ Quái (C18)", cũng có lượt chiếu và đặt vé cao, phản ánh sức hút mạnh mẽ của các bộ phim hành động và kinh dị. Trong khi đó, nhiều phim khác có tần suất rất thấp, dưới 1.000 lượt, cho thấy mức độ quan tâm của khán giả không đồng đều. Sự chênh lệch lớn giữa các phim cho thấy nhu cầu thị trường tập trung vào một số tác phẩm nổi bật, trong khi phần lớn các phim còn lại chỉ thu hút được một lượng nhỏ người xem.

* Tần suất của người bán vé



Hình 2.14: Phân phối tần suất hoạt động của nhân viên thu ngân

Nguồn: Nhóm tác giả tính toán

Nhận xét : Biểu đồ tần suất giao dịch theo nhân viên thu ngân cho thấy sự phân hóa rõ rệt trong khối lượng công việc giữa các nhân viên. emp011 và emp006 là hai nhân viên xử lý nhiều giao dịch nhất, với tần suất hơn 5.000 lượt, cho thấy họ có thể là những người làm việc toàn thời gian hoặc trong các khung giờ cao điểm. Theo sau là các nhân viên như emp012, emp014 và emp005, cũng có lượng giao dịch tương đối cao. Trong khi đó, các nhân viên như emp015, emp013, emp003 có số lượng giao dịch rất thấp, thậm chí thấp hơn cả nền tảng Website, cho thấy họ có thể chỉ làm việc bán thời gian, mới gia nhập, hoặc làm ở khung giờ ít khách. Việc phân bố không đồng đều này có thể phản ánh sự khác biệt về kinh nghiệm, năng suất làm việc hoặc thời gian làm việc của các thu ngân. Đồng thời, việc giao dịch qua Website còn rất hạn chế, cho thấy kênh bán vé trực tuyến chưa được khách hàng ưu tiên sử dụng nhiều và có thể là một cơ hội cải thiện trong tương lai.

## **2.4. Các bước tiền xử lý dữ liệu**

### **2.4.1. Xử lý bảng TICKET**

Trong quá trình kiểm tra dữ liệu bảng TICKET, nhóm đã phát hiện ra sự không khớp giữa tổng giá vé của mỗi đơn hàng (orderid) và trường total được lưu trong bảng. Theo quy định nghiệp vụ, mỗi orderid tương ứng với một đơn đặt vé xem phim và có thể bao gồm nhiều vé khác nhau. Mỗi vé trong đơn sẽ có một mức giá cụ thể (ticket price) và tổng của các mức giá này cần phải bằng với giá trị total thể hiện tổng số tiền của đơn hàng đó.

Tuy nhiên, qua kiểm tra thực tế, nhóm nhận thấy có nhiều trường hợp mà tổng giá vé của các dòng dữ liệu thuộc cùng một orderid không bằng với giá trị total đã ghi nhận. Sai lệch này có thể phát sinh do một số nguyên nhân như: lỗi nhập liệu, sai sót trong quá trình tính toán khi ghi dữ liệu vào hệ thống, hoặc dữ liệu bị cập nhật thiếu đồng bộ giữa bảng chi tiết vé và bảng tổng hợp đơn hàng.

Đây là một vấn đề nghiêm trọng vì nó có thể ảnh hưởng đến độ chính xác trong báo cáo doanh thu, gây ra sai sót trong thống kê tài chính, và ảnh hưởng đến trải nghiệm người dùng nếu có liên quan đến hoàn tiền hoặc khiếu nại. Việc phát hiện và xử lý những sai lệch này là cần thiết để đảm bảo tính nhất quán, toàn vẹn dữ liệu, cũng như củng cố độ tin cậy của hệ thống.

Để khắc phục tình trạng này, nhóm đã sử dụng Python kết hợp với thư viện pandas để rà soát và xử lý dữ liệu. Cụ thể, nhóm thực hiện việc nhóm các dòng dữ liệu theo orderid, tính tổng ticket\_price cho mỗi đơn và so sánh với giá trị total tương ứng. Với những đơn hàng có chênh lệch, nhóm tiến hành fill thêm các dòng bị thiếu, sao cho đảm bảo tổng giá vé khớp với total. Việc này giúp khôi phục dữ liệu thiếu một cách có kiểm soát và hỗ trợ cho các bước phân tích, báo cáo tiếp theo được chính xác và đầy đủ hơn.

|  |
| --- |
| # Giả sử df đã được đọc từ file  # df = pd.read\_csv("data.csv")  # Chỉ giữ lại các cột cần thiết  df\_selected = df[['orderid', 'cashier', 'saledate', 'customerid', 'date', 'time', 'room', 'film', 'ticketcode', 'slot', 'ticket price', 'total', 'popcorn','slot type','ticket type']]  # Tính tổng giá vé theo orderid  total\_ticket\_price = df\_selected.groupby('orderid')['ticket price'].sum().reset\_index()  total\_ticket\_price.rename(columns={'ticket price': 'total\_ticket\_price'}, inplace=True)  # Merge tổng giá vé vào df\_selected  df\_selected = df\_selected.merge(total\_ticket\_price, on='orderid', how='left')  # Tính chênh lệch  df\_selected['difference'] = df\_selected['total'] - df\_selected['total\_ticket\_price']  df\_selected['discount'] = 0 # Khởi tạo cột discount  df\_selected['have\_discount'] = "Không" # Mặc định là "Không"  # Xác định các mức cần thêm dòng mới  adjustment\_values = [45000, 90000, 75000, 160000]  def generate\_unique\_slot(existing\_slots):  while True:  row = random.choice(string.ascii\_uppercase[:8]) # Chọn ngẫu nhiên từ A-H  number = random.randint(1, 17) # Chọn số từ 1 đến 17  new\_slot = f"{row}{number}"  if new\_slot not in existing\_slots:  return new\_slot  # Danh sách chứa dữ liệu cần thêm  new\_rows = []  for \_, row in df\_selected.iterrows():  if row['difference'] in adjustment\_values:  existing\_slots = df\_selected[(df\_selected['date'] == row['date']) &  (df\_selected['time'] == row['time']) &  (df\_selected['room'] == row['room']) &  (df\_selected['film'] == row['film'])]['slot'].tolist()  new\_rows.append({  'orderid': row['orderid'],  'cashier': row['cashier'],  'saledate': row['saledate'],  'customerid': row['customerid'],  'date': row['date'],  'time': row['time'],  'room': row['room'],  'film': row['film'],  'ticket type': row['ticket type'],  'slot type': row['slot type'],  'ticketcode': np.random.randint(10000, 99999),  'slot': generate\_unique\_slot(existing\_slots),  'ticket price': row['difference'],  'total': row['total'],  'popcorn': row['popcorn'],  'total\_ticket\_price': row['total\_ticket\_price'],  'difference': row['difference'],  'discount': 0,  'have\_discount': "Không"  })  elif row['difference'] == 225000:  additional\_tickets = []  remaining\_discount = 225000  while remaining\_discount > 0:  ticket\_price = 75000 if remaining\_discount >= 75000 else 45000  remaining\_discount -= ticket\_price  existing\_slots = df\_selected[(df\_selected['date'] == row['date']) &  (df\_selected['time'] == row['time']) &  (df\_selected['room'] == row['room']) &  (df\_selected['film'] == row['film'])]['slot'].tolist()  additional\_tickets.append({  'orderid': row['orderid'],  'cashier': row['cashier'],  'saledate': row['saledate'],  'customerid': row['customerid'],  'date': row['date'],  'time': row['time'],  'room': row['room'],  'film': row['film'],  'ticketcode': np.random.randint(10000, 99999),  'slot': generate\_unique\_slot(existing\_slots),  'ticket price': ticket\_price,  'total': row['total'],  'popcorn': row['popcorn'],  'total\_ticket\_price': row['total\_ticket\_price'],  'difference': row['difference'],  'discount': 0,  'have\_discount': "Không"  })  new\_rows.extend(additional\_tickets)  elif row['difference'] != 0:  df\_selected.loc[df\_selected['orderid'] == row['orderid'], 'discount'] = row['difference']  df\_selected.loc[df\_selected['orderid'] == row['orderid'], 'have\_discount'] = "Có"  # Tạo DataFrame từ danh sách mới và thêm vào df\_selected  if new\_rows:  df\_new = pd.DataFrame(new\_rows)  df\_selected = pd.concat([df\_selected, df\_new], ignore\_index=True) |

* Mục tiêu :
* Xác định các đơn hàng có sự chênh lệch giữa tổng giá trị vé và tổng tiền thanh toán.
* Bổ sung dòng dữ liệu bị thiếu tương ứng với giá trị chênh lệch, dựa trên các mức giá vé chuẩn.
* Gắn nhãn cho các đơn hàng có chiết khấu thay vì lỗi dữ liệu.
* Quy trình xử lý :

Đầu tiên, nhóm giữ lại các cột cần thiết trong bảng dữ liệu bằng cách lọc qua df\_selected. Sau đó, nhóm tính tổng ticket price cho mỗi orderid bằng lệnh groupby và gộp (merge) trở lại bảng chính để so sánh với trường total. Một cột mới tên là difference được tạo ra, đại diện cho độ chênh giữa giá trị total và tổng các dòng ticket price thực tế. Tiếp theo, nhóm xử lý các dòng có chênh lệch cụ thể nằm trong danh sách [45000, 90000, 75000, 160000] — đây là các mức giá vé phổ biến. Với mỗi dòng như vậy, nhóm tự động tạo thêm một dòng dữ liệu mới mô phỏng một vé bị thiếu, trong đó:

* ticket price được gán bằng giá trị chênh lệch.
* slot được tạo ngẫu nhiên nhưng đảm bảo không trùng với slot đã có.
* ticketcode là số ngẫu nhiên để mô phỏng mã vé.

Đối với những đơn có chênh lệch lớn hơn như 225000, nhóm chia nhỏ chênh lệch này thành nhiều vé có giá trị 75000 hoặc 45000, rồi tạo nhiều dòng tương ứng để bổ sung. Trường hợp chênh lệch không thuộc các mức trên, nhóm giả định rằng đơn hàng có thể đã áp dụng chiết khấu. Khi đó, nhóm không tạo dòng mới mà cập nhật cột discount và gán nhãn have\_discount là "Có" để ghi nhận tình trạng. Cuối cùng, nếu có dòng mới được tạo, nhóm kết hợp (concat) chúng với bảng dữ liệu gốc df\_selected để tạo ra một bảng hoàn chỉnh, phục vụ cho các bước phân tích tiếp theo.

Sau khi hoàn tất bước kiểm tra và xử lý các đơn hàng có orderid hợp lệ, bước tiếp theo nhóm tập trung xử lý các giao dịch được thực hiện qua kênh website. Đặc thù của nhóm dữ liệu này là nhiều dòng bị thiếu giá trị orderid, gây khó khăn cho việc tổng hợp và phân tích theo đơn hàng.

Để khắc phục, nhóm đã áp dụng một chiến lược gán mã đơn hàng tự động. Cụ thể, nhóm sử dụng công thức định dạng MAK (viết tắt của "Mã Khách") kết hợp với một số thứ tự được sinh tự động để tạo ra các orderid mới. Tuy nhiên, để đảm bảo tính hợp lý và nhất quán, nếu hai dòng dữ liệu có cùng customerid và được thực hiện trong cùng một ngày (date), nhóm sẽ gán cùng một orderid cho cả hai dòng. Cách làm này giúp tái cấu trúc dữ liệu thành các đơn hàng hoàn chỉnh, phản ánh chính xác hành vi mua vé của khách hàng, đồng thời hỗ trợ các bước xử lý tiếp theo như tính doanh thu, phân tích hành vi và tối ưu hóa hệ thống.

# **CHƯƠNG 3: ỨNG DỤNG CÁC PHƯƠNG PHÁP VÀO BÀI TOÁN KINH DOANH**

#### **3.1 Bài toán và mục tiêu mô hình**

Dựa trên dữ liệu thu thập được từ hoạt động bán vé trong tháng 5/2019 của rạp chiếu phim, nhóm nghiên cứu nhận thấy rằng dữ liệu không chỉ đơn thuần phản ánh doanh thu hay số lượng vé, mà còn chứa đựng nhiều thông tin giá trị về hành vi tiêu dùng và đặc điểm của khách hàng. Do đó, bài toán mà nhóm đề xuất tập trung giải quyết là phân tích hành vi mua vé của khách hàng nhằm phân loại các nhóm khách có đặc điểm tương đồng, từ đó hỗ trợ doanh nghiệp trong việc đưa ra các quyết định kinh doanh chính xác và cá nhân hóa dịch vụ tốt hơn.

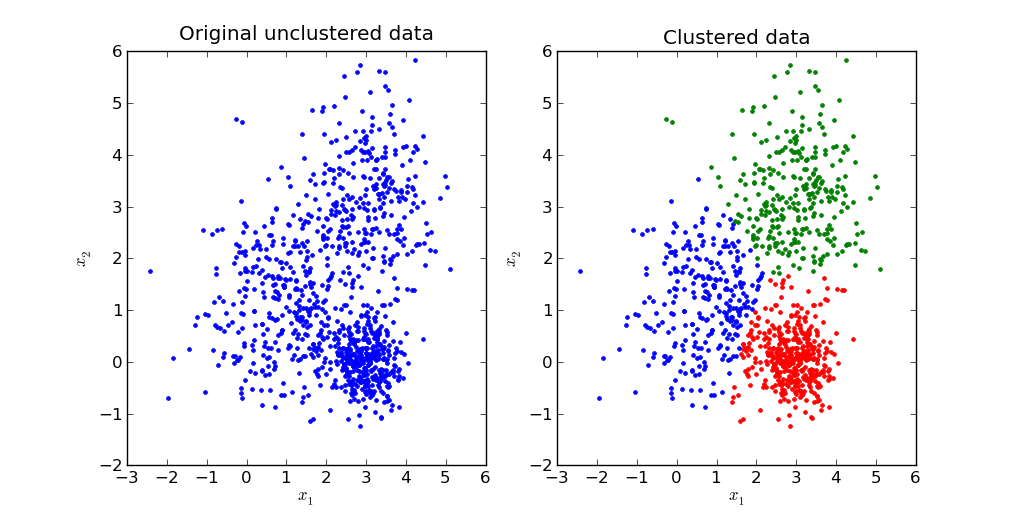
Cụ thể, nhóm sử dụng hai thuật toán phân cụm là K-Means và DBSCAN để thực hiện việc phân cụm khách hàng, từ đó xây dựng các chương trình ưu đãi, chiến dịch marketing hay gói sản phẩm phù hợp hơn với từng nhóm cụ thể. Sau khi áp dụng cả hai thuật toán, nhóm sẽ tiến hành phân tích và so sánh hai kết quả phân cụm cũng như lựa chọn ra mô hình phù hợp với thực tế doanh nghiệp để ứng dụng cho các phân tích trong tương lai.

## **3.2. Lựa chọn phương pháp**

Đối với bài toán phân cụm khách hàng thì phương pháp phân cụm KMeans được nhóm cân nhắc chọn đầu tiên do tính đơn giản và khả năng mở rộng phù hợp với kích thước của tập dữ liệu. Bên cạnh đó, phương pháp này hoạt động dựa trên việc phân chia dữ liệu vào những nhóm đã được cố định trước. Điều này phù hợp với việc nhóm đã có những nghiên cứu trước đó về việc phân cụm khách hàng thành các phân khúc nào và kết quả của việc phân cụm cũng dễ diễn giải để triển khai nhanh chóng.

Tuy nhiên, nhóm chúng em cũng cần phải thận trọng vì có thể những gì mà nhóm nghiên cứu trước về việc phân cụm khách hàng có thể không phù hợp với thực tế dữ liệu của doanh nghiệp. Vì vậy nhóm đã bổ sung thêm thuật toán DBSCAN nhằm khắc phục điểm yếu là phải xác định số cụm từ trước của thuật toán KMeans. Thuật toán DBSCAN có thể giúp nhóm xác định được các nhóm khách hàng ngách hoặc là các hành vi bất thường trong dữ liệu.

**3.3. Mô tả chi tiết các mô hình sử dụng**

****

Hình 3.1: Minh họa cho mô hình phân cụm

*Nguồn: Mubaris*

Để giải quyết bài toán phân cụm khách hàng thì nhóm chúng em đã sử dụng hai mô hình phân cụm khác nhau là KMeans và DBSCAN đại diện cho hai hướng tiếp cận khác nhau trong mô hình học máy không giám sát.

Đối với mô hình KMeans hoạt động theo cơ chế chia dữ liệu thành một số cụm cố định bằng cách khởi tạo ngẫu nhiên các tâm cụm sau đó lặp lại quá trình gán mỗi điểm dữ liệu vào cụm có khoảng cách gần nhất với tâm cụm và cập nhật lại tâm cho đến khi đạt trạng thái hội tụ. Các siêu tham số quan trọng bao gồm số lượng cụm (n\_clusters), phương pháp khởi tạo tâm cụm (init), số lần khởi tạo (n\_init), số vòng lặp tối đa (max\_iter) và giá trị khởi tạo ngẫu nhiên (random\_state) nhằm đảm bảo tính nhất quán của kết quả. Có thể xử lý dữ liệu lớn với thời gian tính toán tương đối nhanh, đặc biệt khi số cụm nhỏ. Hoạt động tốt khi các cụm có hình dạng cầu và tách biệt rõ ràng. Tuy nhiên, thuật toán này có một số nhược điểm là ta phải biết hoặc đoán được số lượng cụm phù hợp trước khi chạy và con số này không phải lúc nào cũng dễ dàng thấy được. Kết quả có thể khác nhau nếu chọn tâm cụm ban đầu khác nhau và điều này có thể dẫn đến tối ưu cục bộ. Dữ liệu nhiễu hoặc điểm ngoại lệ có thể kéo lệch tâm cụm, ảnh hưởng đến kết quả phân cụm. Hàm tính toán khoảng cách trong K-Means là hàm Euclid, nên không phù hợp với dữ liệu không số hóa hoặc dữ liệu phân loại gốc.

Mô hình DBSCAN là mô hình phân cụm dựa trên mật độ không yêu cầu xác định trước số cụm hoạt động bằng cách xác định điểm lõi có đủ số lượng điểm lân cận trong bán kính eps và mở rộng cụm từ các điểm này cho đến khi không thể mở rộng thêm. Trong mô hình này thì các siêu tham số đáng chú ý của DBSCAN bao gồm eps khoảng cách tối đa để hai điểm được xem là lân cận (eps), số điểm tối thiểu để xác định điểm lõi (min\_samples) cùng với công thức dùng để tính khoảng cách (euclid, manhattan,..). Có thể phân cụm tốt với các hình dạng phức tạp: cụm hình trăng lưỡi liềm, hình chữ C, v.v, Có khả năng nhận diện điểm nhiễu (noise/outliers) và không gộp chúng vào bất kỳ cụm nào. Nếu các cụm có mật độ khác nhau, DBSCAN có thể gộp chúng lại hoặc chia cụm sai. Hiệu năng giảm đáng kể khi dữ liệu có nhiều chiều do khoảng cách Euclidean mất hiệu lực.

Cả hai mô hình được triển khai bằng ngôn ngữ Python sử dụng thư viện scikit-learn cho các thuật toán phân cụm. Bên cạnh đó, nhóm cũng sử dụng các thư viện như pandas để đọc dữ liệu dạng bảng và numpy để hỗ trợ các phép toán số học và kiểm soát tham số trong quá trình huấn luyện và đánh giá mô hình.

## **3.4. Quy trình huấn luyện mô hình**

Trước khi huấn luyện mô hình thì nhóm chúng em sẽ phải tiến hành xử lý dữ liệu. Đầu tiên là nhóm sẽ loại bỏ các cột không liên quan và xử lý giá trị thiếu, chuyển đổi định dạng ngày sinh thành thuộc tính tuổi, xử lý giá trị bất thường. Tiếp theo nhóm sẽ tính các chỉ số xác định hành vi của khách hàng theo mô hình RFM (Recency - Frequency - Monetary). Đối với Recency thì nhóm sẽ xác định khách hàng đã mua vé gần đây không bằng cách lấy ngày giao dịch gần nhất của họ và tính số ngày kể từ lần mua gần nhất đến bây giờ. Còn với Frequency thì nhóm sẽ đếm số vé từng khách hàng đã mua (dùng customerid). Monetary sẽ được tính là tổng số tiền khách hàng đã chi. Tiếp theo là mã hóa biến phân loại. Trong các biến phân loại thì với biến ‘Job’ và ‘Industry’ thì nhóm sử dụng phương pháp one-hot encoding còn giới tính thì sử dụng cat.codes để mã hóa cột phân loại 'gender'. Sau khi đã xử lý xong các dữ liệu thì nhóm sẽ chuẩn hóa biến liên tục như ‘Recency', 'Frequency', 'Monetary', 'age' và các thể loại phim bằng StandardScaler.

Sau khi đã xử lý xong dữ liệu thì nhóm sẽ bắt đầu ứng dụng mô hình KMeans vào để giải quyết bài toán. Bởi vì đối với mô hình KMeans thì số cụm (n\_cluster) là siêu tham số quan trọng tác động đến kết quả của mô hình. Vì vậy, nhóm đã phân tích Elbow Method dựa trên WCSS nhằm để xác định n\_clusters tối ưu.

Đối với mô hình DBSCAN thì nhóm sẽ tập trung trong việc hiệu chỉnh siêu tham số eps và min\_samples. Chất lượng phân cụm được đánh giá bằng chỉ số Silhouette Score nhằm đo lường mức độ tách biệt và tính chặt chẽ giữa các cụm. Do đó, quá trình hiệu chỉnh siêu tham số được thực hiện bằng giá trị eps trong khoảng từ 0.3 đến 1.5 với bước nhảy 0.1 và giá trị min\_samples từ 3 đến 9 mỗi cặp tham số được kiểm tra và tính toán silhouette score để tìm ra bộ tham số mang lại giá trị silhouette score cao nhất. Tuy nhiên cần lưu ý là trong quá trình hiệu chỉnh siêu tham số thì các trường hợp dữ liệu chỉ có một cụm hoặc toàn bộ bị gán nhãn nhiễu được loại trừ để đảm bảo phân cụm có ý nghĩa sau quá trình thử nghiệm. Sau cùng, nhóm sẽ ghi nhận best\_score cùng best\_params tương ứng với mỗi bước thử nghiệm để áp dụng bộ tham số tối ưu.

# **CHƯƠNG 4: KẾT QUẢ MÔ HÌNH**

## **4.1. Tiền xử lý dữ liệu**

Dữ liệu được xử lý từ nhiều nguồn khác nhau, bao gồm thông tin về đơn vé (TICKET), khách hàng (CUSTOMER) và phim (LISTEDIN). Đầu tiên, dữ liệu từ các tệp Excel được đọc vào, sau đó ngày giao dịch trong bảng đơn vé được chuẩn hóa sang định dạng thời gian để xác định ngày phân tích là ngày giao dịch mới nhất. Sau đó, để tránh đếm trùng đơn hàng, chỉ giữ lại các đơn hàng duy nhất theo orderid. Từ đó, các chỉ số RFM (Recency, Frequency, Monetary) được tính cho từng khách hàng: số ngày kể từ giao dịch gần nhất, số lần giao dịch và tổng chi tiêu.

Tên phim trong bảng vé và bảng liệt kê phim được chuẩn hóa (viết thường, loại bỏ khoảng trắng) để phục vụ việc nối dữ liệu. Sau khi nối, thông tin thể loại phim (từ 6 cột thể loại khác nhau) được gom lại, chuẩn hóa và ánh xạ về các nhóm thể loại chính như sci-fi, action, comedy, v.v. Dữ liệu sau đó được tách riêng từng thể loại phim trong mỗi vé thành các dòng riêng biệt, nhờ vậy có thể dễ dàng đếm được số lần mỗi khách hàng xem từng thể loại phim cụ thể.

Tiếp theo, thời gian xem phim được phân loại thành các khung giờ trong ngày (Sáng, Trưa, Chiều, Tối, Khuya) dựa trên giờ giao dịch. Dữ liệu được nhóm lại theo customerid để đếm số vé mua theo từng khung giờ. Tương tự, vùng chỗ ngồi (dựa trên ký tự đầu tiên của ‘slot’) cũng được trích xuất và tổng hợp theo khách hàng.

Cuối cùng, tất cả các thông tin được tổng hợp lại theo từng customerid, bao gồm chỉ số RFM, số lần xem từng thể loại phim, hành vi theo thời gian và khu vực chỗ ngồi. Bảng dữ liệu cuối cùng được kết hợp với thông tin khách hàng từ bảng CUSTOMER để tạo thành tập dữ liệu hoàn chỉnh phục vụ cho mô hình phân cụm khách hàng.

## **4.2. Kết quả các mô hình**

Sau khi đọc dữ liệu kết quả của tiền xử lý phía trước, các cột không cần thiết như A đến K (chỗ ngồi) và các khung thời gian trong ngày đã được loại bỏ để tinh gọn dữ liệu. Các giá trị khuyết trong các cột ‘DOB’, giới tính và nghề nghiệp được điền bằng giá trị phổ biến nhất nhằm đảm bảo tính đồng nhất, trong khi các giá trị thiếu ở cột ngành nghề (industry) được thay thế bằng chuỗi 'Nan'. Bên cạnh đó, các thông tin không liên quan như địa chỉ và website cũng bị loại bỏ. Cột ngày sinh (DOB) sau đó được chuyển đổi về định dạng ngày chuẩn bằng cách xử lý cả trường hợp định dạng số (Excel serial date) và chuỗi, rồi được chuyển thành cột tuổi (age). Các giá trị tuổi bất hợp lý như ≤ 0 được thay thế bằng giá trị tuổi phổ biến nhất. Cột ‘DOB’ sau khi chuyển đổi thành tuổi cũng bị loại bỏ để tránh dư thừa.

Tiếp theo, dữ liệu dạng phân loại như nghề nghiệp và ngành nghề được mã hóa bằng phương pháp One-Hot Encoding, trong khi giới tính được mã hóa dạng số nguyên sử dụng .cat.codes. Dữ liệu sau đó được chuẩn hóa bằng StandardScaler đối với các cột số như Recency, Frequency, Monetary, tuổi và sở thích thể loại phim, nhằm đảm bảo các biến có cùng đơn vị và tỷ lệ khi đưa vào mô hình phân cụm.

Ảnh có chứa văn bản, ảnh chụp màn hình, hàng, biểu đồ

Nội dung do AI tạo ra có thể không chính xác.

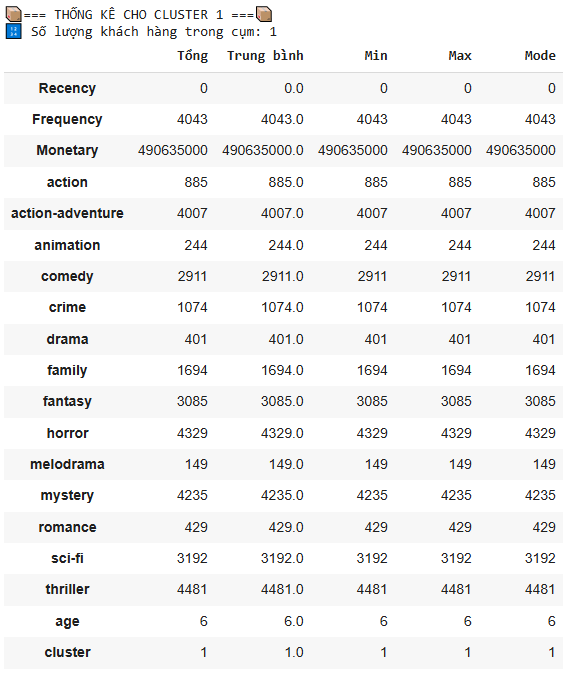
Hình 4.1: Biểu đồ Elbow – Chọn số cụm tối ưu

*Nguồn: Nhóm tác giả tính toán*

Để xác định số cụm tối ưu, phương pháp Elbow được sử dụng bằng cách tính tổng bình phương sai trong cụm (WCSS) với số cụm từ 1 đến 10, từ đó giúp lựa chọn giá trị k phù hợp dựa trên điểm gãy của đồ thị. Tuy nhiên, Elbow cho ra kết quả 2 cụm là tốt nhất, không phù hợp với quy mô và mục đích bài toán.

Nhóm quyết định thử chọn 4 là số cụm cho bài toán này. Mô hình KMeans được áp dụng với số cụm mặc định là 4 và kết quả phân cụm được lưu vào cột cluster. Kết quả phân cụm cuối cùng được gộp lại với dữ liệu gốc để phục vụ cho phân tích chuyên sâu từng nhóm khách hàng như thống kê trung bình, tổng, giá trị nhỏ nhất, lớn nhất và mode cho từng cụm.

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

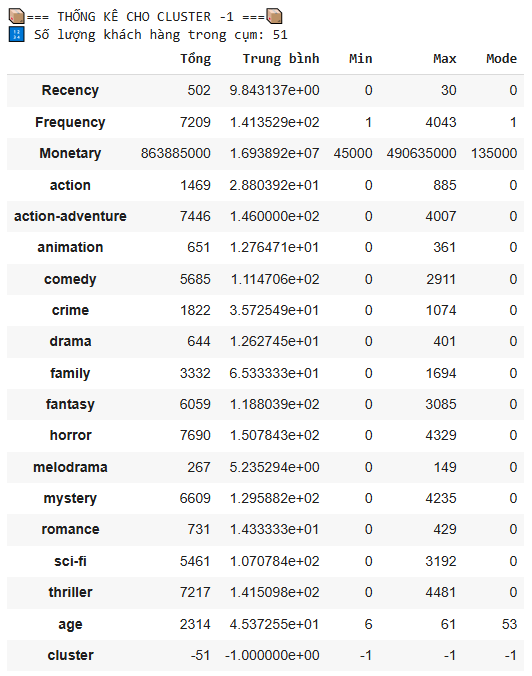
Nội dung do AI tạo ra có thể không chính xác.Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình 4.2: Mô tả kết quả phân cụm KMeans

*Nguồn: Nhóm tác giả tính toán*

Ở phần tiếp theo, thuật toán DBSCAN được triển khai nhằm khai thác khả năng phát hiện các cụm có hình dạng phức tạp cũng như loại bỏ các điểm ngoại lai (outlier). Sau khi loại bỏ cột customerid và nhãn cụm từ KMeans, quy trình dò tìm lưới được tiến hành với phạm vi eps từ 0.3 đến 1.4 và min\_samples từ 3 đến 9 nhằm tìm ra tổ hợp thông số tối ưu. Với mỗi tổ hợp, nếu số lượng cụm nằm trong khoảng hợp lệ từ 2 đến 10, điểm silhouette score sẽ được tính để đánh giá mức độ phân tách giữa các cụm. Dựa trên kết quả dò tìm, tổ hợp siêu tham số tối ưu được xác định là eps bằng 1.4 và min\_samples bằng 5, với số cụm được tạo là 3 và đạt điểm silhouette score là 0.1472. Đây là tổ hợp mang lại hiệu quả phân tách tốt nhất trong các lựa chọn đã thử. Mô hình DBSCAN sau đó được áp dụng với các tham số eps và min\_samples đã lựa chọn, và kết quả phân cụm thu được được hợp nhất trở lại với dữ liệu ban đầu. Từ đó, quá trình phân tích thống kê mô tả từng cụm khách hàng được thực hiện tương tự như với phương pháp KMeans, nhằm hiểu rõ hơn về đặc điểm và hành vi của từng nhóm khách hàng.

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.Ảnh có chứa văn bản, ảnh chụp màn hình, số, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình 4.3: Mô tả kết quả phân cụm DBSCAN

*Nguồn: Nhóm tác giả tính toán*

## **4.3. Phân tích và so sánh**

### **4.3.1. Phân tích kết quả từ thuật toán KMean**

Sau khi tiến hành phân cụm dữ liệu khách hàng của rạp chiếu phim, ta thu được 4 nhóm chính với những đặc điểm hành vi khác biệt rõ rệt. Cụm đầu tiên gồm phần lớn khách hàng (2.230 người), là nhóm có tần suất đến rạp thấp, chi tiêu rất ít và đã lâu không quay lại. Họ thường là người trẻ (khoảng 27 tuổi), từng xem một số thể loại như hành động, viễn tưởng, trinh thám và phiêu lưu, nhưng nhìn chung mức độ tương tác rất thấp. Đây là nhóm khách hàng "ngủ quên", có khả năng từng đến rạp vài lần rồi rơi vào trạng thái không hoạt động. Đối với nhóm này, rạp có thể triển khai các chiến dịch kích hoạt lại như gửi mã giảm giá, gợi ý phim thuộc thể loại họ từng yêu thích hoặc nhắc nhở qua email/phần mềm.

Ngược lại, cụm thứ hai chỉ có một khách hàng duy nhất nhưng có hành vi nổi bật. Người này vừa mới đến rạp (recency bằng 0), xem phim hơn 4.000 lần và chi tiêu gần 5 tỷ đồng. Tuy nhiên, độ tuổi được ghi nhận chỉ là 6 tuổi, điều này có thể là lỗi dữ liệu hoặc tài khoản dùng chung. Dù vậy, đây là khách hàng cực kỳ giá trị và nên được chăm sóc theo hướng VIP, như tặng vé miễn phí, ưu đãi cá nhân hóa hoặc mời tham gia chương trình khách hàng thân thiết. Cụm này cho thấy rằng dữ liệu cần được kiểm tra lại để đảm bảo tính chính xác.

Cụm thứ ba cũng chỉ có một khách hàng, có hành vi tương tự cụm thứ hai: rất thường xuyên đến rạp (gần 2.900 lần), vừa mới tương tác và chi tiêu khoảng 3,3 tỷ đồng. Người này ưa chuộng các thể loại như kinh dị, hành động, phiêu lưu và trinh thám. Mặc dù số lượng khách trong cụm chỉ là 1, nhưng đây cũng là nhóm khách đặc biệt cần được ghi nhận và phục vụ theo chế độ riêng biệt.

Cuối cùng, cụm thứ tư là nhóm khách hàng chiếm số lượng lớn thứ hai với hơn 2.200 người. Nhóm này có tần suất ghé rạp ở mức trung bình (khoảng 2 lần), chi tiêu ổn định (trung bình hơn 200.000 đồng) và quay lại rạp khá gần đây (6 ngày trước). Họ thường ở độ tuổi gần 29, yêu thích các thể loại như hài, kinh dị, phiêu lưu và trinh thám. Đây là nhóm khách hàng hoạt động tích cực và là đối tượng tiềm năng để gia tăng giá trị vòng đời. Rạp nên tập trung upsell nhóm này bằng các combo đồ ăn, thẻ thành viên, ưu đãi định kỳ, đồng thời khuyến khích họ giới thiệu bạn bè để mở rộng tập khách hàng trung thành.

Tóm lại, việc phân cụm này giúp rạp chiếu phim hiểu rõ đặc điểm từng nhóm khách hàng để từ đó triển khai các chiến lược marketing và chăm sóc phù hợp. Thay vì áp dụng một chiến dịch chung, rạp có thể cá nhân hoá nội dung quảng bá, ưu đãi và dịch vụ theo từng nhóm khách cụ thể – điều này sẽ giúp tối ưu chi phí, tăng doanh thu và giữ chân khách hàng hiệu quả hơn.

### **4.3.2. Phân tích kết quả từ thuật toán DBSCAN**

Trong cụm đầu tiên (tương ứng với cluster -1, bị DBSCAN đánh dấu là nhiễu), có 51 khách hàng với tần suất mua hàng trung bình lên đến hơn 141 lần, và chi tiêu trung bình đạt khoảng 1,69 triệu – cao vượt trội so với các cụm khác. Đây là nhóm khách hàng có mức độ tương tác rất cao, thường xuyên quay lại mua sắm, và có sở thích rõ rệt đối với các thể loại như phim kinh dị, hành động, phiêu lưu, và giật gân. Recency (khoảng thời gian kể từ lần mua gần nhất) của nhóm này cũng tương đối thấp, cho thấy họ hoạt động gần đây. Mặc dù bị coi là nhiễu, thực chất đây là nhóm khách hàng có giá trị rất lớn, xứng đáng được ưu tiên chăm sóc đặc biệt.

Cụm thứ hai (tương ứng với cluster 0) là nhóm khách hàng đông nhất, với hơn 3300 người. Tuy nhiên, đây lại là nhóm có tần suất mua hàng và chi tiêu trung bình rất thấp – lần lượt chỉ khoảng 1,6 và 17.000. Họ cũng không có sở thích nổi bật với bất kỳ thể loại phim nào, cho thấy mức độ gắn kết kém. Đây là nhóm khách hàng đại trà, ít tương tác, cần các chiến dịch marketing kích hoạt lại hoặc khảo sát thêm để hiểu rõ hành vi.

Cụm thứ ba (tương ứng với cluster 1) có hơn 1000 khách hàng, với tần suất mua hàng và chi tiêu trung bình nhỉnh hơn cụm thứ hai, khoảng 1,7 và hơn 212.000. Họ thể hiện sự quan tâm đến các thể loại như kinh dị, phiêu lưu, và thần thoại, có thể từng là nhóm khách hàng trung thành nhưng đã bắt đầu "nguội lạnh". Độ tuổi trung bình của nhóm này khoảng 23, nên cũng có thể là nhóm khách hàng trẻ, năng động.

Cuối cùng, cụm thứ tư (cluster 2) chỉ gồm 27 người nhưng lại có chi tiêu khá cao (trung bình hơn 211.000) và hứng thú với các thể loại phiêu lưu, hành động và kinh dị. Tuy nhiên, độ tuổi trung bình lại chỉ khoảng 7,3 – có thể đây là lỗi dữ liệu hoặc phản ánh nhóm trẻ em được người lớn mua giúp.

Tổng thể, kết quả phân cụm bằng DBSCAN giúp phát hiện rõ một nhóm khách hàng đặc biệt giá trị (cụm đầu tiên) mà thuật toán KMeans không thể hiện rõ. Nếu mục tiêu là tìm kiếm nhóm khách hàng VIP để chăm sóc chuyên biệt và cá nhân hóa trải nghiệm, kết quả từ DBSCAN sẽ phù hợp hơn. Ngược lại, nếu muốn chia nhóm ổn định và đồng đều để triển khai các chiến dịch marketing đại trà hoặc phân tích tổng thể, kết quả từ KMeans sẽ dễ sử dụng hơn.

## **4.4. Dashboard tổng hợp kết quả:**

Trong bối cảnh chuyển đổi số mạnh mẽ hiện nay, Dashboard đóng vai trò then chốt trong việc trực quan hóa dữ liệu, hỗ trợ các nhà quản trị đưa ra quyết định nhanh chóng và chính xác dựa trên cơ sở dữ liệu thực tiễn. Đối với lĩnh vực kinh doanh rạp chiếu phim, Dashboard không chỉ cung cấp cái nhìn toàn diện về hoạt động vận hành mà còn giúp khai thác sâu các hành vi khách hàng, tối ưu hóa doanh thu và nâng cao hiệu quả dịch vụ.

Dashboard tháng 5/2019 của rạp chiếu phim được xây dựng trên nền tảng Power BI, một công cụ Business Intelligence mạnh mẽ cho phép tổng hợp, trực quan hóa và phân tích dữ liệu một cách trực quan, dễ hiểu. Dashboard này bao gồm các thành phần chính như:

* Các chỉ số KPI (Key Performance Indicators): Tổng doanh thu, số lượng vé bán, số suất chiếu, tỷ lệ đặt vé qua website, tỷ lệ tiêu thụ bắp rang. Các chỉ số này giúp quản lý nhanh chóng nắm bắt tình hình kinh doanh hiện tại.
* Biểu đồ động (Interactive Charts): Phân tích doanh thu, lượng vé bán theo ngày, khung giờ, thể loại phim, nhóm nghề nghiệp khách hàng. Các biểu đồ được trình bày đa dạng dưới dạng cột, đường, tròn, thanh ngang nhằm làm nổi bật các mối tương quan và xu hướng.
* Slicer (Bộ lọc dữ liệu động): Cho phép người dùng tương tác trực tiếp với dữ liệu, lọc theo khung giờ, thể loại phim, nhóm khách hàng, từ đó hỗ trợ các phân tích chuyên sâu và linh hoạt.

Thông qua Dashboard, nhà quản trị có thể theo dõi sát sao biến động doanh thu, nhận diện các khung giờ vàng, xác định nhóm khách hàng mục tiêu, cũng như phát hiện các điểm yếu trong hoạt động như tỷ lệ mua vé online thấp, hiệu suất F&B chưa tối ưu. Điều này giúp nâng cao hiệu quả vận hành, hỗ trợ ra quyết định nhanh, kịp thời, đồng thời làm cơ sở để thiết lập các chiến lược marketing, điều chỉnh lịch chiếu, phát triển dịch vụ bổ trợ phù hợp với hành vi và nhu cầu khách hàng.

### **4.3.1. Overview Dashboard**

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình 4.4: Dashboard Overview - Tổng quát về tình trạng doanh nghiệp

*Nguồn: Nhóm tác giả tính toán*

#### **4.3.1.1. Các chỉ số đánh giá tổng quan về rạp phim:**

Trong dashboard tổng quan tháng 5/2019, bốn KPI trọng yếu được thể hiện gồm: tổng số vé bán ra (Total Tickets), tỷ lệ đặt vé qua website (Website Rate), tỷ lệ tiêu thụ bắp rang (Popcorn Rate), và tổng số suất chiếu (Screenings). Việc phân tích cụ thể từng chỉ số không chỉ phản ánh hiệu quả hiện tại mà còn là cơ sở để đưa ra các khuyến nghị cải tiến trong tương lai.

* Tổng doanh thu đạt 1,688 triệu đồng là một chỉ số phản ánh doanh thu gộp từ việc bán vé xem phim trong tháng 5 năm 2019 của rạp chiếu phim. Đây là chỉ số cốt lõi để đánh giá hiệu quả khai thác nội dung điện ảnh, mức độ thu hút của các bộ phim được trình chiếu, và hành vi tiêu dùng giải trí của khách hàng trong kỳ báo cáo. Doanh thu này không bao gồm các khoản thu phụ trợ như bán bắp nước, quảng cáo, hay các dịch vụ bổ sung khác. Trong bối cảnh tháng 5 thường trùng với mùa cao điểm (do học sinh bắt đầu nghỉ hè), mức doanh thu này được đánh giá là trung bình – khá, chưa phản ánh sự bùng nổ vượt trội. Tuy tổng doanh thu ở mức khá, nhưng nếu phân tách các chỉ số liên quan (Popcorn Rate chỉ đạt 0.35%, Website Rate 0.27%), có thể thấy doanh thu chủ yếu vẫn đến từ bán vé trực tiếp tại rạp, trong khi các mảng doanh thu tiềm năng như dịch vụ đi kèm, nền tảng số chưa được khai thác hiệu quả. Điều này cho thấy cơ cấu doanh thu chưa cân bằng, phụ thuộc nhiều vào nguồn thu truyền thống, tiềm ẩn rủi ro trong dài hạn.
* Tổng số vé bán ra đạt 36,000 trong tháng 5/2019, tương đương với trung bình khoảng 1,200 vé mỗi ngày. Đây là một con số đáng ghi nhận đối với một rạp có quy mô vừa, cho thấy mức độ thu hút khách hàng ổn định trong suốt tháng. Tuy nhiên, nếu xét đến yếu tố mùa vụ (tháng 5 trùng kỳ nghỉ hè đầu mùa), con số này vẫn chưa phản ánh mức tăng trưởng mạnh.
* Tỷ lệ đặt vé qua website - chỉ số này ở mức cực kỳ thấp, cho thấy chưa đến 1% khách hàng mua vé thông qua nền tảng trực tuyến. Điều này phản ánh rằng khách hàng vẫn chủ yếu mua vé trực tiếp tại quầy, hoặc ứng dụng/website đặt vé hiện tại chưa tối ưu, chưa thân thiện hoặc chưa đủ hấp dẫn. Trong bối cảnh xu hướng chuyển đổi số và thói quen tiêu dùng tiện lợi đang lên ngôi, việc tỷ lệ đặt vé online thấp là một điểm yếu rõ ràng. Nếu không cải thiện, rạp sẽ bỏ lỡ cơ hội gia tăng hiệu suất đặt vé, giảm tải cho quầy vé và thu thập dữ liệu hành vi người dùng cho các chiến dịch marketing mục tiêu.
* Tỷ lệ tiêu thụ bắp rang - đây là một trong những chỉ số quan trọng phản ánh mức độ hiệu quả của mảng doanh thu phụ trợ (ancillary revenue) – thường chiếm tỉ trọng lớn trong tổng lợi nhuận của rạp chiếu phim. Với tỷ lệ chỉ đạt 0.35%, điều này cho thấy rạp đang gặp khó khăn trong việc thuyết phục khách hàng sử dụng sản phẩm F&B đi kèm. Có thể lý giải bằng nhiều nguyên nhân: mức giá sản phẩm chưa hợp lý, thiếu các combo khuyến mãi, vị trí quầy bán chưa tối ưu, hoặc nhân viên chưa thực hiện tốt upselling. Đây là một cảnh báo cần cải thiện ngay để khai thác hiệu quả hơn nguồn thu quan trọng này.
* Tổng số suất chiếutổng cộng trong tháng 5 có 733 suất chiếu, tương đương trung bình 24 suất/ngày. Đây là mức khai thác tương đối hợp lý nếu rạp sở hữu từ 3–5 phòng chiếu. Tuy nhiên, nếu so sánh với tổng số vé bán ra thì trung bình mỗi suất chiếu chỉ đạt khoảng 49 khách – một con số chưa cao. Điều này cho thấy có thể có hiện tượng phân bổ suất chiếu chưa tối ưu, lãng phí khung giờ thấp điểm hoặc chiếu các phim có sức hút kém trong các giờ vàng. Việc tối ưu lịch chiếu, tập trung vào khung giờ cao điểm và tăng suất cho các phim hút khách sẽ giúp cải thiện hiệu quả vận hành đáng kể.

#### **4.3.1.2. Phân tích doanh thu và số suất chiếu theo từng ngày trong tháng**

* Biểu đồ **“**Revenue and Screenings by Date of Month” thể hiện mối quan hệ giữa tổng doanh thu từ vé xem phim (Sum of ticket price) và số suất chiếu phim mỗi ngày (Show Count) trong suốt tháng 5/2019. Trục tung bên trái biểu diễn giá trị doanh thu (tính bằng đồng Việt Nam), trong khi trục tung bên phải biểu thị số lượng suất chiếu mỗi ngày.
* Phân tích xu hướng chính:
  + Biến động đầu tháng (ngày 1–10): Giai đoạn đầu tháng ghi nhận sự dao động mạnh mẽ cả về doanh thu và số suất chiếu. Đáng chú ý là vào khoảng ngày 3 và ngày 4 ghi nhận mức doanh thu và số suất chiếu giảm mạnh đột biến so với các ngày còn lại trong tháng. Nguyên nhân được xác định là do quốc tang, dẫn đến việc tạm dừng hoặc hạn chế các hoạt động giải trí, bao gồm cả việc chiếu phim tại rạp. Điều này phản ánh đúng thực tế hoạt động văn hóa trong nước, và là yếu tố ngoại cảnh ảnh hưởng rõ rệt đến hiệu suất kinh doanh.. Sau đó, cả hai chỉ số tăng trở lại vào ngày 5 và đạt đỉnh vào khoảng ngày 8–10, cho thấy hiệu quả tối ưu khi suất chiếu và nội dung được điều chỉnh hợp lý.
  + Giai đoạn giữa tháng (ngày 11–20): Doanh thu và số suất chiếu bắt đầu có xu hướng tách biệt: trong khi số lượng suất chiếu duy trì ổn định ở mức cao, thì doanh thu lại giảm sút trong một số ngày (đặc biệt là khoảng ngày 15–17). Điều này phản ánh khả năng chất lượng phim hoặc thời lượng phim không hấp dẫn, dẫn đến doanh thu trên mỗi suất chiếu không cao. Rạp có thể đã mở nhiều suất chiếu nhưng không đạt được lượng khách tương ứng.
  + Giai đoạn cuối tháng (ngày 21–31): Xu hướng ổn định trở lại, với số suất chiếu duy trì ổn định trên mức 20 suất/ngày. Doanh thu có những biến động nhẹ, với các điểm tăng đột biến vào các ngày 25 và 30, nhiều khả năng rơi vào cuối tuần hoặc giai đoạn phim mới ra rạp. Đây là minh chứng cho thấy hiệu ứng ngày nghỉ và phim bom tấn mới phát hành vẫn giữ vai trò quan trọng trong việc tăng trưởng doanh thu cục bộ.
* Từ biểu đồ, có thể thấy rằng có sự tương quan tương đối chặt chẽ giữa số lượng suất chiếu và doanh thu theo ngày. Những ngày có số suất chiếu cao thường đi kèm với doanh thu tăng, tuy nhiên không phải lúc nào cũng hoàn toàn tỷ lệ thuận. Do đó, cần:
  + Tối ưu hóa lịch chiếu linh hoạt theo ngày trong tháng, đặc biệt ưu tiên các ngày cuối tuần hoặc lễ để phát hành phim mới.
  + Giám sát sát sao hiệu quả suất chiếu theo từng ngày để có điều chỉnh nhanh chóng với các phim không đạt doanh thu kỳ vọng.
  + Kết hợp khuyến mãi và combo vé vào các ngày doanh thu thấp trong tuần nhằm kích thích nhu cầu và tăng lấp đầy ghế trống.

#### **4.3.1.3. Phân tích doanh thu và suất chiếu theo thời gian trong ngày**

* Biểu đồ trên thể hiện sự phân bổ giữa tổng doanh thu từ giá vé và số lượng suất chiếu theo bốn khung giờ chính trong ngày: Sáng – Chiều – Tối – Khuya. Trục tung bên trái biểu diễn tổng doanh thu, trong khi trục tung bên phải biểu thị số lượng suất chiếu tương ứng.
* Phân tích chi tiết:
  + Khung giờ tối: Đây là khung thời gian có doanh thu cao nhất, đạt khoảng 0.8 tỉ đồng, đồng thời cũng là khung giờ có số suất chiếu cao thứ hai (gần 250 suất). Điều này phản ánh rõ ràng rằng khán giả có xu hướng đến rạp vào buổi tối – sau khi kết thúc công việc và học tập. Đây cũng là khoảng thời gian thuận tiện cho các nhóm đối tượng chủ lực như teenager và white collar, những người có khả năng chi trả cao và thời gian rảnh buổi tối. Việc doanh thu cao dù không phải là khung giờ có nhiều suất chiếu nhất cho thấy hiệu quả khai thác suất chiếu tối là rất tốt.
  + Khung giờ chiều: Chiều là khung giờ có số suất chiếu nhiều nhất, vượt mốc 300 suất, nhưng doanh thu lại thấp hơn so với buổi tối. Điều này cho thấy có dấu hiệu dư thừa suất chiếu, đặc biệt nếu nhiều suất chiếu không đạt tỷ lệ lấp đầy ghế. Khung giờ chiều thường phục vụ student và một phần white collar nghỉ ca sớm, tuy nhiên hiệu suất sinh lời trên mỗi suất chiếu cần được xem xét lại. Việc tái cơ cấu số lượng suất chiếu chiều sang tối có thể giúp tối ưu hóa chi phí vận hành và cải thiện tỷ suất lợi nhuận.
  + Khung giờ sáng: Với doanh thu ở mức thấp và số suất chiếu dưới 100, đây là khung giờ có hoạt động tương đối hạn chế. Tuy nhiên, sáng lại là khoảng thời gian phù hợp cho các nhóm như blue collar hoặc khán giả cao tuổi. Nếu có điều chỉnh để phục vụ các bộ phim ngắn – dễ tiếp cận – giá vé thấp thì hoàn toàn có thể khai thác thêm tiềm năng từ khung giờ này.
  + Khung giờ khuya: Đây là khung giờ có doanh thu và suất chiếu thấp nhất, gần như không đóng góp nhiều cho tổng thể. Dữ liệu phản ánh rằng hoạt động chiếu phim khuya không hiệu quả, đồng thời có thể làm gia tăng chi phí vận hành (nhân sự, điện năng, bảo trì…). Do vậy, nên hạn chế hoặc cắt giảm hoàn toàn các suất chiếu sau 22h, trừ khi có sự kiện đặc biệt hoặc phát hành phim hot.
* Đề xuất cho các chiến lược:
  + Tối ưu hóa khung giờ chiều và tối, chuyển một phần suất chiếu từ chiều sang tối để đạt tỷ suất doanh thu tốt hơn.
  + Rà soát suất chiếu sáng và khuya, chỉ nên duy trì với phim có nhu cầu rõ ràng từ nhóm đối tượng phù hợp hoặc sự kiện đặc biệt.
  + Xây dựng chính sách giá vé linh hoạt theo giờ, ví dụ khuyến mãi vé sáng hoặc combo buổi chiều, để nâng cao hiệu quả lấp đầy ghế.

4.3.1.4. Xu hướng sản phẩm phụ trợ (bỏng ngô) dựa trên thời lượng phim:

* Biểu đồ **“**Count of popcorn by duration type” phản ánh số lượng bỏng ngô được tiêu thụ theo các nhóm thời lượng phim khác nhau, bao gồm: khoảng 1 tiếng rưỡi, khoảng 2 tiếng, và khoảng 3 tiếng. Đây là một chỉ số quan trọng nhằm xác định mối liên hệ giữa độ dài của bộ phim và hành vi tiêu dùng sản phẩm phụ trợ – cụ thể là bỏng ngô – từ đó đưa ra đề xuất tối ưu hóa dịch vụ bán hàng và trải nghiệm khách hàng.
* Phân tích dữ liệu
  + Phim dài khoảng 2 tiếng có số lượng bỏng ngô được bán ra nhiều nhất, chiếm tỷ trọng lớn nhất trong ba nhóm thời lượng. Điều này cho thấy rằng khi thời lượng phim kéo dài ở mức trung bình – không quá ngắn cũng không quá dài – người xem có xu hướng tiêu thụ nhiều đồ ăn vặt hơn để duy trì trải nghiệm giải trí liên tục. Đồng thời, thời lượng này cũng phổ biến trong các phim bom tấn hành động hoặc phim dành cho giới trẻ (teenager), nhóm khách hàng tiêu thụ bỏng ngô cao nhất.
  + Phim dài khoảng 1 tiếng rưỡi ghi nhận mức tiêu thụ bỏng ngô ở mức trung bình. Đây là mức thời lượng ngắn hơn nhưng vẫn đủ lâu để tạo nhu cầu tiêu dùng đồ ăn trong rạp. Đối tượng khán giả chính có thể là student và white collar – những người có thời gian giải trí giới hạn nhưng vẫn duy trì thói quen mua bỏng ngô như một phần trải nghiệm.
  + Phim có thời lượng khoảng 3 tiếng ghi nhận lượng tiêu thụ bỏng ngô thấp nhất. Nguyên nhân có thể đến từ việc các bộ phim quá dài thường khiến người xem e ngại tiêu thụ nhiều thực phẩm, tránh ảnh hưởng đến sự tập trung hoặc phải rời rạp giữa chừng. Ngoài ra, số lượng phim có độ dài này cũng khá ít trong lịch chiếu tháng 5 nên cơ hội tiêu thụ sản phẩm kèm theo cũng bị hạn chế.
* Dữ liệu từ biểu đồ cho thấy rằng thời lượng phim có mối liên hệ trực tiếp đến hành vi tiêu dùng bỏng ngô, và thời lượng khoảng 2 tiếng là điểm tối ưu để kích thích doanh thu từ sản phẩm phụ trợ. Rạp chiếu phim có thể tận dụng thông tin này để:
  + Thiết kế các combo bỏng ngô phù hợp với từng loại phim, ví dụ như combo tiêu chuẩn cho phim 2 tiếng, mini combo cho phim ngắn và ưu đãi refill cho phim dài.
  + Tăng chiến dịch tiếp thị bỏng ngô và các sản phẩm phụ trợ ở các suất chiếu phim dài vừa (2 tiếng), đặc biệt vào cuối tuần.
  + Khuyến khích khán giả mua trước combo trực tuyến khi đặt vé đối với các phim bom tấn có thời lượng lý tưởng để tăng tỷ lệ chuyển đổi.

#### **4.3.1.5. Doanh thu theo nhóm đối tượng nghề nghiệp**

* Biểu đồ tròn “Ticket Revenue Distribution by Occupation” phản ánh tỷ trọng doanh thu bán vé theo từng nhóm nghề nghiệp của khách hàng đến rạp phim trong tháng 5 năm 2019. Đây là một chỉ số quan trọng để xác định nhóm khách hàng mục tiêu và từ đó đề xuất các chiến lược marketing và lịch chiếu phù hợp.
* Phân tích chi tiết các nhóm khách hàng:
  + Teenager (Tuổi vị thành niên): Chiếm tỷ trọng lớn nhất với 64.24% tổng doanh thu (tương đương khoảng 1 tỷ đồng), đây là nhóm khách hàng chủ lực của rạp phim trong giai đoạn khảo sát. Điều này phản ánh xu hướng tiêu dùng giải trí cao ở nhóm tuổi trẻ, năng động, đặc biệt là vào thời điểm cuối tuần hoặc sau giờ học. Sự thống trị của nhóm này gợi ý rằng rạp cần tiếp tục đầu tư vào các bộ phim phù hợp với thị hiếu tuổi teen như hành động, viễn tưởng hoặc hoạt hình có yếu tố hài hước, cùng với các chương trình khuyến mãi nhóm hoặc giảm giá học sinh.
  + Student (Sinh viên): Đứng thứ hai với 13.34%, nhóm sinh viên cũng là đối tượng khách hàng tiềm năng do có thói quen đi xem phim vào chiều tối trong tuần hoặc cuối tuần. Tuy nhiên, tỷ lệ doanh thu còn thấp hơn khá nhiều so với teenager, có thể do hạn chế về tài chính hoặc lịch học. Việc điều chỉnh giá vé sinh viên hoặc mở rộng các khung giờ chiếu phim ngắn, giá rẻ có thể giúp tăng sức hút với nhóm này.
  + White Collar (Nhân viên văn phòng): Chiếm 7.65% doanh thu, nhóm nhân viên văn phòng là nhóm khách có mức thu nhập ổn định, tuy nhiên thời gian giải trí bị giới hạn. Đây là nhóm khách có tiềm năng cao nếu rạp điều chỉnh lịch chiếu phù hợp – chẳng hạn như tăng suất chiếu vào buổi tối các ngày trong tuần và các suất đặc biệt vào cuối tuần.
  + Blue Collar (Lao động phổ thông): Góp phần 7.63% vào doanh thu, nhóm lao động phổ thông có đặc thù thời gian làm việc cố định và ít linh hoạt hơn. Do đó, các phim có thời lượng ngắn và khung giờ phù hợp như buổi sáng hoặc đầu chiều nên được ưu tiên để thu hút nhóm này.
  + Specialist (Chuyên gia): Nhóm chuyên gia đóng góp rất ít vào doanh thu (0.02%), gần như không đáng kể. Có thể nhóm này ưu tiên các hoạt động giải trí khác hoặc yêu cầu về nội dung phim cao hơn (phim nghệ thuật, phim quốc tế). Rạp có thể cân nhắc khai thác nhóm này thông qua các chương trình chiếu phim chuyên đề hoặc sự kiện đặc biệt.
* Với biểu đồ phân tích theo nghề nghiệp cho thấy nhóm teenager, student và white collar là ba phân khúc chủ lực mà rạp chiếu phim cần tập trung. Việc xây dựng các chiến lược tiếp thị và tối ưu lịch chiếu phù hợp với thói quen tiêu dùng và thời gian rảnh của từng nhóm sẽ giúp tăng trưởng doanh thu hiệu quả hơn.

#### **4.3.1.6. Top 5 Phim theo Số Lượng Vé Bán và Số Suất Chiếu**

* Biểu đồ “Top 5 Movies by Ticket Sales and Show Count” cho thấy rõ sự khác biệt về hiệu quả khai thác giữa các bộ phim trong tháng 5 năm 2019. Trong đó, hai chỉ số được phân tích gồm số vé bán ra và số suất chiếu – là cơ sở quan trọng để đánh giá hiệu suất kinh doanh của từng bộ phim.
  + MẸ MA THAN KHÓC LA LLORONA đứng đầu bảng về cả lượng vé bán và số suất chiếu, cho thấy đây là bộ phim chủ lực của rạp trong tháng. Việc đạt gần 10.000 vé bán ra với số suất chiếu cao cho thấy khả năng thu hút khán giả tốt và có thể đến từ chiến lược quảng bá mạnh hoặc nội dung hấp dẫn.
  + AVENGERS: HỒI KẾT có số vé bán ra gần bằng MA nhưng số suất chiếu lại thấp hơn đáng kể. Điều này thể hiện tỷ lệ lấp đầy phòng chiếu cao, tức là hiệu suất doanh thu trên mỗi suất chiếu vượt trội. Đây là phim có thương hiệu toàn cầu, cho thấy rạp nên ưu tiên giữ suất chiếu dù tần suất không quá nhiều.
  + NGÔI ĐỀN KỲ LẠ (B) và JOHN WICK 3: CUỘC CHIẾN TRÊN THÁNH ĐỊA đều có mức vé bán và suất chiếu ở mức trung bình. Điều này phản ánh việc rạp đang phân phối suất chiếu dàn trải, nhưng hiệu suất chưa đạt mức tối ưu.
  + POKÉMON: THÁM TỬ PIKACHU tuy có thương hiệu quốc tế nhưng lại xếp cuối trong top 5 cả về doanh thu và suất chiếu. Điều này cho thấy bộ phim chưa phù hợp thị hiếu nội địa thời điểm đó, hoặc gặp sự cạnh tranh mạnh từ các phim khác cùng thời điểm.
* Từ biểu đồ này, có thể đưa ra đề xuất tối ưu hóa lịch chiếu dựa trên hiệu quả doanh thu như sau:
  + Ưu tiên phân bổ nhiều suất chiếu hơn cho các phim có tỷ lệ lấp đầy cao, như Avengers: Hồi Kết, để tối đa hóa doanh thu trên mỗi suất.
  + Giảm hoặc tái cấu trúc suất chiếu của các phim có hiệu suất thấp hơn như Pokémon, nhằm tránh lãng phí tài nguyên phòng chiếu.
  + Đối với các phim có hiệu suất trung bình như John Wick hoặc Ngôi Đền Kỳ Lạ, có thể tăng suất chiếu vào các khung giờ thấp điểm để khai thác thêm nhu cầu tiềm năng mà không ảnh hưởng đến doanh thu khung giờ cao điểm.

#### **4.3.1.7. Xu hướng thể loại phim:**

* Biểu đồ “Total Revenue of Movies Genres” thể hiện doanh thu gộp theo từng thể loại phim được trình chiếu trong tháng 5 năm 2019. Dữ liệu này có vai trò quan trọng trong việc đánh giá mức độ ưa chuộng nội dung và định hướng chiến lược chọn phim đầu tư trong tương lai. Trục hoành biểu thị doanh thu, trục tung là tên thể loại.
* Trong quá trình phân tích doanh thu theo thể loại, một bộ phim có thể mang nhiều thể loại khác nhau (ví dụ: Action – Fantasy – Sci-fi). Tuy nhiên, hệ thống dữ liệu không thể gán toàn bộ doanh thu của phim đó cho tất cả thể loại một cách đồng đều vì sẽ gây ra sai lệch thống kê. Do đó, để phản ánh chính xác hơn mức độ đóng góp của từng thể loại vào tổng doanh thu, một hệ thống trọng số đã được áp dụng dựa trên vị trí xuất hiện và mức độ đại diện của từng thể loại trong mô tả phim. Cách tính cụ thể:
  + Thể loại chính (đứng đầu): nhận 60% tổng doanh thu của bộ phim.
  + Thể loại phụ: được phân bổ phần còn lại (40%) tùy theo số lượng thể loại phụ đi kèm. Cụ thể:

|  |  |  |
| --- | --- | --- |
| **Tổng số thể loại** | **Tỷ lệ thể loại chính** | **Tỷ lệ các thể loại phụ** |
| 2 (1 chính + 1 phụ) | 60% | 40% |
| 3 (1 chính + 2 phụ) | 60% | 25% – 15% |
| 4 (1 chính + 3 phụ) | 60% | 20% – 15% – 5% |
| 5 (1 chính + 4 phụ) | 60% | 17.5% – 12.5% – 7.5% – 2.5% |
| 6 (1 chính + 5 phụ) | 60% | 15% – 10% – 7.5% – 5% – 2.5% |

*Nguồn: Nhóm tác giả tổng hợp*

* Thể loại dẫn đầu doanh thu là Action-Adventure và Horror xếp thứ hai, sát nút với thể loại hành động. Hai thể loại này có mức doanh thu vượt trội, cho thấy đây là hai dòng phim chủ lực đóng góp phần lớn cho tổng doanh thu 1,688 triệu đồng. Comedy (hài) xếp thứ ba, nhưng doanh thu cách biệt đáng kể so với 2 thể loại đầu. Điều này cho thấy mức độ quan tâm có nhưng chưa đủ mạnh để cạnh tranh về mặt doanh thu. Các thể loại Fantasy, Mystery, Thriller có doanh thu trung bình, cho thấy mức tiêu thụ ổn định nhưng không nổi bật. Các thể loại như Science fiction (khoa học viễn tưởng) hoặc thể loại Action ở cuối bảng có doanh thu rất thấp, dưới mức 0.1 tỷ VND. Điều này phản ánh có thể là vì các phim thuộc thể loại này ít được chiếu trong tháng, nội dung phim không hấp dẫn với đối tượng khách hàng mục tiêu tại rạp, cạnh tranh kém so với các bom tấn hành động và kinh dị. Việc thể loại hành động và kinh dị dẫn đầu phù hợp với xu hướng tiêu dùng phổ biến của giới trẻ, vốn là đối tượng khách hàng chủ lực của rạp phim. Đây là những thể loại:
  + Có nhịp phim nhanh, kỹ xảo cao.
  + Trải nghiệm tốt trên màn ảnh lớn (tối ưu hóa trải nghiệm rạp).
  + Được hỗ trợ truyền thông mạnh từ các hãng phim quốc tế.

Dựa trên kết quả của biểu đồ, các đề xuất chiến lược cho rạp phim:

* + Tập trung vào phim hành động và kinh dị trong các dịp cao điểm, đầu tư thêm vào lịch chiếu, khung giờ vàng, và truyền thông.
  + Định vị rõ nhóm khán giả tiềm năng cho các thể loại trung bình (như hài, fantasy) để tối ưu cách tiếp cận.
  + Cân nhắc đa dạng hóa nội dung để tránh quá lệ thuộc vào 1–2 dòng phim.
  + Kiểm tra và chuẩn hóa dữ liệu thể loại trong hệ thống bán vé, tránh trùng lặp thể loại gây sai số phân tích.

### **4.3.2. Customer Dashboard**

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Phông chữ

Nội dung do AI tạo ra có thể không chính xác.

Hình 4.5: Customer Dashboard - Chân dung khách hàng

*Nguồn: Nhóm tác giả tổng hợp*

Nhìn chung, trong tháng 5/2019, rạp chiếu phim đã bán tổng cộng 35,508 vé, với số lượng vé trung bình trên mỗi đơn hàng đạt 7.93 và doanh thu trung bình trên mỗi giao dịch khoảng 115,000. Đáng chú ý, tỷ lệ khách mua vé theo hình thức cặp đôi cao đáng kể, lên đến 51%, cho thấy đây hiện đang là nhóm khách hàng đóng vai trò chủ đạo trong tổng doanh thu.

Xét theo hình thức đi xem phim, nhóm khách đi theo cặp đôi mang lại doanh thu cao nhất, tiếp theo là nhóm khách đi theo nhóm. Trong khi đó, nhóm khách đi một mình có tỷ lệ doanh thu thấp hơn rõ rệt. Điều này có thể phản ánh rằng các chiến lược hướng đến trải nghiệm cho cặp đôi khá hiệu quả. Đây là tín hiệu tích cực để doanh nghiệp tiếp tục phát triển các mô hình dịch vụ đặc thù cho cặp đôi.

Về độ tuổi, khách hàng từ 18–25 tuổi vẫn là nhóm chiếm ưu thế vượt trội cả về doanh thu lẫn số lượng người dùng, với gần 15,000 khách hàng trong tháng. Nhóm 26–30 tuổi xếp thứ hai nhưng cách biệt đáng kể, trong khi các nhóm tuổi trên 30 có tỷ trọng rất thấp. Điều này củng cố nhận định rằng nhóm khách hàng mục tiêu hiện tại là giới trẻ — đặc biệt là sinh viên hoặc người đi làm trẻ — những người có xu hướng giải trí thường xuyên và dễ tiếp cận qua nền tảng số.

Phân tích theo mức chi tiêu, khách hàng chi dưới 200,000 vẫn chiếm số lượng đông đảo nhất, kế tiếp là nhóm chi trên 400,000, trong khi nhóm chi tiêu trung bình (200,000–400,000) lại có số lượng thấp hơn đáng kể. Đây là một điểm thú vị, cho thấy có hai cực rõ ràng trong hành vi tiêu dùng: một bên là khách hàng tiết kiệm, một bên là khách hàng có khả năng chi tiêu cao. Điều này mở ra hai hướng chiến lược: (1) duy trì các combo giá rẻ, vé ưu đãi cho nhóm đại chúng, và (2) phát triển thêm các dịch vụ cao cấp như phòng chiếu VIP, ghế đôi, hoặc dịch vụ theo yêu cầu cho nhóm khách hàng chi mạnh tay.

Lượng vé bán ra trong tháng dao động theo từng ngày, với những đỉnh cao rơi vào những ngày cuối tuần. Gần như tất cả các ngày trong tháng đều có số lượng vé bán ra trên 500 vé. Duy chỉ có ngày 3 bán được 154 vé và ngày 4 không bán vé. Bởi vì đây là khoảng thời gian diễn ra sự kiện Quốc tang nguyên Chủ tịch nước Lê Đức Anh và các hoạt động vui chơi giải trí phải ngừng lại. Rạp chiếu phim có thể khai thác các khoảng thời gian thấp điểm để tăng cường quảng bá, triển khai ưu đãi nhằm tối ưu công suất phòng chiếu.

### **4.3.3. Scheduling Dashboard**

Ảnh có chứa văn bản, ảnh chụp màn hình, Song song, thiết kế

Nội dung do AI tạo ra có thể không chính xác.

Hình 4.6: Scheduling Dashboard - Xây dựng lịch chiếu

*Nguồn: Nhóm tác giả tổng hợp*

#### **4.3.3.1. Xu hướng mua vé theo ngày trong tuần**

Biểu đồ"Ticket Sales by Day of the Week" cho thấy rõ ràng rằng nhóm tuổi teen (teenager) là đối tượng có tần suất mua vé cao nhất, đặc biệt vào cuối tuần (Thứ bảy và Chủ nhật). Các nhóm như học sinh, sinh viên (student) và nhân viên văn phòng (white collar) cũng ghi nhận doanh số cao hơn vào hai ngày này, điều này phản ánh thói quen giải trí vào thời điểm rảnh rỗi. Ngược lại, các nhóm lao động phổ thông (blue collar) và chuyên gia (specialist) có xu hướng mua vé phân bổ đồng đều hơn trong tuần, cho thấy mức độ linh hoạt về thời gian của họ.

#### **4.3.3.2. Khung giờ xem phim ưa thích**

Biểu đồ "Ticket Sales by Industry and Time Frame"giúp làm rõ thời điểm trong ngày mà các nhóm đối tượng thường xuyên đến rạp. Nhóm teenager chiếm ưu thế rõ rệt vào các khung giờ buổi tối (Tối) và khuya (Khuya), thể hiện sự ưu tiên cho các hoạt động giải trí vào thời điểm muộn. Student và white collar có xu hướng xem phim vào buổi chiều (Chiều) và tối, phù hợp với lịch học và làm việc cố định. Trong khi đó, blue collar và specialist có xu hướng tránh các suất chiếu muộn, chủ yếu tập trung vào buổi sáng (Sáng) và chiều, phù hợp với lịch trình lao động sớm và cần nghỉ ngơi vào buổi tối.

#### **4.3.3.3. Thời lượng phim và mức độ ưa thích**

Qua biểu đồ "Average Movie Watching Time by Industry Group", có thể thấy nhóm teenager có mức độ chấp nhận các bộ phim có thời lượng dài hơn (khoảng 2 tiếng) cao nhất, vượt trội so với các nhóm còn lại. Ngược lại, các nhóm như student, white collar và blue collar có xu hướng ưu tiên các bộ phim có thời lượng ngắn hơn (khoảng 1–1.5 tiếng), phù hợp với nhu cầu giải trí ngắn gọn và không ảnh hưởng đến công việc hoặc học tập.

#### **4.3.3.4. Loại phim được ưa chuộng**

Biểu đồ **"Movie Rating Rankings by Industry Group"** cho thấy phim C18 – thường là phim hành động, tâm lý người lớn – thu hút đông đảo lượng khán giả, đặc biệt là từ nhóm teenager và white collar. Trong khi đó, phim C13 và P – thường là thể loại nhẹ nhàng, hài hước hoặc gia đình – được student và teenager ưa chuộng. Phim C16 có lượt xem thấp, đặt ra câu hỏi về mức độ hấp dẫn và khả năng thu hút khán giả của thể loại này.

#### **4.3.3.5. Đề xuất tối ưu hóa lịch chiếu**

* Tối ưu lịch chiếu theo thời gian:

Để tối ưu lịch chiếu và đáp ứng nhu cầu giải trí của từng nhóm đối tượng, cần có sự điều chỉnh phù hợp về thời gian và loại phim. Cụ thể, nên tăng số suất chiếu các phim C18 và C13 vào buổi tối và cuối tuần, nhằm phục vụ tốt hơn cho hai nhóm khán giả chủ lực là teenager và white collar, vốn có xu hướng xem phim nhiều vào khung giờ này. Bên cạnh đó, việc bổ sung các suất chiếu vào chiều và tối trong các ngày thường sẽ giúp tiếp cận hiệu quả hơn với nhóm student, những người có thời gian rảnh chủ yếu sau giờ học. Đồng thời, cần cân nhắc tổ chức thêm các suất chiếu vào buổi sáng và chiều trong tuần dành cho nhóm blue collar, đặc biệt là với các phim có thời lượng ngắn (khoảng 1 đến 1.5 tiếng) để phù hợp với lịch trình làm việc và nghỉ ngơi của họ. Những điều chỉnh này không chỉ góp phần tăng lượng vé bán ra mà còn nâng cao trải nghiệm khách hàng thông qua sự linh hoạt và phù hợp trong lịch chiếu.

* Phân loại phim & thời lượng:

Việc phân bổ thời lượng phim cũng cần được cân nhắc kỹ lưỡng dựa trên thói quen và sở thích của từng nhóm khán giả. Theo dữ liệu từ dashboard, nhóm teenager có xu hướng xem phim dài nhiều nhất, do đó nên ưu tiên các bộ phim có thời lượng khoảng 2 tiếng khi phục vụ đối tượng này, đặc biệt vào buổi tối và cuối tuần. Trong khi đó, student và white collar lại thể hiện sự ưa chuộng đối với những bộ phim có thời lượng ngắn hơn (dưới 2 tiếng). Đối với hai nhóm này, việc lên lịch chiếu các phim ngắn vào giữa tuần – khi thời gian rảnh bị giới hạn bởi công việc và học tập – sẽ giúp nâng cao khả năng tiếp cận và tăng hiệu quả khai thác suất chiếu trong các khung giờ thấp điểm.

* Phân phối suất chiếu theo ngày trong tuần:

Lịch chiếu vào từng thời điểm trong tuần cũng cần được thiết kế linh hoạt để phù hợp với nhu cầu thực tế và tối ưu hóa chi phí vận hành. Cuối tuần (Thứ Bảy và Chủ Nhật) là thời điểm lượng khán giả đến rạp cao nhất, do đó nên ưu tiên các suất chiếu "hot" với những bộ phim có đánh giá cao như C18 và C13, nhằm thu hút tối đa nhóm đối tượng chính là teenager và white collar. Ngược lại, vào các ngày giữa tuần, khi nhu cầu giải trí có xu hướng giảm, rạp phim có thể giảm số lượng suất chiếu hoặc lựa chọn các bộ phim phổ thông, có thời lượng ngắn để phục vụ nhóm khán giả có thời gian hạn chế, đồng thời giảm thiểu chi phí vận hành mà vẫn đảm bảo hiệu quả khai thác trong những khung giờ thấp điểm.

# 

# **PHỤ LỤC**

## **Phụ lục 1: Khám phá dữ liệu ban đầu (EDA)**

* [EDA\_BI\_FOR\_CUSTOMER](https://colab.research.google.com/drive/1pFgsrf0LRB_OSghHhk_tyrUfVYEIolG7?usp=sharing)
* [EDA\_BI\_FOR\_TICKET](https://colab.research.google.com/drive/1SVW09UWNnq66zN9naR7PO3sJI7EiJUCF?usp=sharing)

## **Phụ lục 2: Tiền xử lý dữ liệu**

* [Ticket\_processing](https://colab.research.google.com/drive/1mtn7pCPV6eRcxDeJrkeV2yyzy1Wj9hTx?usp=sharing)